Génération automatique de questionnaires à choix multiples pédagogiques : évaluation de l'homogénéité des options

Van-Minh Pho

UNIVERSITÉ PARIS-SUD

ÉCOLE DOCTORALE 427 :
INFORMATIQUE PARIS SUD

Laboratoire : *LIMSI-CNRS*

INFORMATIQUE

par

Van-Minh PHO

Génération automatique de questionnaires à choix multiples pédagogiques : évaluation de l'homogénéité des options

Date de soutenance : 24/09/2015

Composition du jury :

Directeur de thèse :	Brigitte GRAU	**Professeur** (ENSIIE, LIMSI-CNRS)
Co-directeur de thèse :	Anne-Laure LIGOZAT	**Maître de conférences** (ENSIIE, LIMSI-CNRS)
Rapporteurs :	Thierry POIBEAU	**Directeur de recherche CNRS** (Lattice)
	Cédrick FAIRON	**Professeur** (Université Catholique de Louvain, CENTAL)
Examinateurs :	Nathalie GUIN	**Maître de conférences HDR** (Université Lyon 1, LIRIS)
	Sophie ROSSET	**Directrice de recherche CNRS** (LIMSI-CNRS)
Membre invitée :	Yolaine BOURDA	**Professeur** (SUPELEC, LRI)

RÉSUMÉ

Ces dernières années ont connu un renouveau des Environnements Informatiques pour l'Apprentissage Humain. Afin que ces environnements soient largement utilisés par les enseignants et les apprenants, ils doivent fournir des moyens pour assister les enseignants dans leur tâche de génération d'exercices. Parmi ces exercices, les Questionnaires à Choix Multiples (QCM) sont très présents. Cependant, la rédaction d'items à choix multiples évaluant correctement le niveau d'apprentissage des apprenants est une tâche complexe. Des consignes ont été développées pour rédiger manuellement des items, mais une évaluation automatique de la qualité des items constituerait un outil pratique pour les enseignants.

Nous nous sommes intéressé à l'évaluation automatique de la qualité des distracteurs (mauvais choix de réponse). Pour cela, nous avons étudié les caractéristiques des distracteurs pertinents à partir de consignes de rédaction de QCM. Cette étude nous a conduit à considérer que l'homogénéité des distracteurs et de la réponse est un critère important pour valider les distracteurs. L'homogénéité est d'ordre syntaxique et sémantique. Nous avons validé la définition de l'homogénéité par une analyse de corpus de QCM, et nous avons proposé des méthodes de reconnaissance automatique de l'homogénéité syntaxique et sémantique à partir de cette analyse.

Nous nous sommes ensuite focalisé sur l'homogénéité sémantique des distracteurs. Pour l'estimer automatiquement, nous avons proposé un modèle d'ordonnancement par apprentissage, combinant différentes mesures d'homogénéité sémantique. L'évaluation du modèle a montré que notre méthode est plus efficace que les travaux existants pour estimer l'homogénéité sémantique des distracteurs.

ABSTRACT

Recent years have seen a revival of Intelligent Tutoring Systems. In order to make these systems widely usable by teachers and learners, they have to provide means to assist teachers in their task of exercise generation. Among these exercises, multiple-choice tests are very common. However, writing Multiple-Choice Questions (MCQ) that correctly assess a learner's level is a complex task. Guidelines were developed to manually write MCQs, but an automatic evaluation of MCQ quality would be a useful tool for teachers.

We are interested in automatic evaluation of distractor (wrong answer choice) quality. To do this, we studied characteristics of relevant distractors from multiple-choice test writing guidelines. This study led us to assume that homogeneity between distractors and answer is an important criterion to validate distractors. Homogeneity is both syntactic and semantic. We validated the definition of homogeneity by a MCQ corpus analysis, and we proposed methods for automatic recognition of syntactic and semantic homogeneity based on this analysis.

Then, we focused our work on distractor semantic homogeneity. To automatically estimate it, we proposed a ranking model by machine learning, combining different semantic homogeneity measures. The evaluation of the model showed that our method is more efficient than existing work to estimate distractor semantic homogeneity.

REMERCIEMENTS

Je souhaite remercier les personnes qui ont permis de mener à bien ma thèse :

Anne-Laure Ligozat et Brigitte Grau pour leur encadrement, leurs encouragements, leur soutien, et surtout, leur patience et leur disponibilité. Je les remercie également pour les relectures (rapides !) de nos articles et du manuscrit de thèse.

Yolaine Bourda pour m'avoir conseillé et guidé, notamment dans le domaine de l'EIAH.

Cédrick Fairon et Thierry Poibeau pour avoir accepté de rapporter mon manuscrit de thèse et pour les remarques détaillées et pertinentes qui ont suivi leurs relectures.

Sophie Rosset et Nathalie Guin pour avoir accepté d'examiner mes travaux et soulevé des pistes de recherches de par leurs questions et discussions.

Thomas François et Cédrick Fairon pour leur accueil au sein du Cental où j'y ai effectué un séjour de recherche, leur disponibilité et leurs conseils. Je remercie particulièrement Thomas François pour m'avoir entraîné au badminton.

Gabriel Illouz pour les conseils prodigués tout au long de ma thèse, qui m'ont été utiles dans le cadre de ma recherche et de mon activité d'enseignement.

Les membres du LIMSI, et particulièrement du groupe ILES, pour m'avoir permis de travailler dans des conditions idéales, ainsi que pour les pauses, les parties de loups-garous et de badminton, ainsi que les pots.

Les membres du Cental, étendus aux participants des Funky Fridays, pour leur accueil chalereux à Louvain-la-Neuve.

Mes amis, mes camarades (et également amis) du 10, et ma famille, qui m'ont soutenu continuellement et qui m'ont supporté depuis plusieurs années.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

1 INTRODUCTION

Ces dernières années ont connu un renouveau des Environnements Informatiques pour l'Apprentissage Humain (EIAH). Ces environnements informatiques ont pour objectifs de susciter, favoriser, accompagner et personnaliser l'apprentissage humain et sont utilisés dans des situations d'interaction présentielle ou à distance. Les EIAH couvrent différents types d'outils destinés aux apprenants comme les environnements de simulation, les jeux sérieux et les plates-formes de formation à distance et d'apprentissage collaboratif. Certains EIAH sont également destinés à l'assistance à l'enseignant, notamment pour rédiger des exercices à destination des apprenants.

Parmi ces EIAH, les plates-formes de Formation en Ligne Ouvertes à Tous (*MOOC, Massive Open Online Course*) ont eu un grand succès. Ces plates-formes permettent la formation d'un nombre non limité d'apprenants dispersés géographiquement et n'effectuent pas de sélection préalable des apprenants. Dans le cadre de ces formations, les enseignants fournissent des ressources pédagogiques, tels que des cours et des exercices. Ces exercices permettent à l'enseignant d'évaluer les apprenants afin de vérifier leur niveau d'acquisition de connaissances dans une thématique donnée. Ils permettent également aux apprenants de s'auto-évaluer afin de vérifier s'ils ont assimilé une notion qu'ils sont censés maîtriser.

Dans le cadre de ces plates-formes, les enseignants proposent principalement des Questionnaires à Choix Multiples (QCM). En effet, ces questionnaires sont constitués d'items pour lesquels l'apprenant se voit proposer un choix parmi des options de réponses préétablies. La correction de tels questionnaires est très simple car il s'agit seulement de vérifier la correspondance entre les choix sélectionnés par l'apprenant et les réponses à ces items. La correction de QCM peut être automatisée, ce qui facilite l'évaluation d'un nombre non limité d'apprenants, et favorise l'auto-évaluation de connaissances par ces apprenants.

Bien que les QCM soient largement utilisés, ils peuvent présenter des défauts de conception qui peuvent engendrer un biais dans l'évaluation des apprenants. En effet, si les QCM donnent des indices sur la réponse correcte, ou sont trop compliqués, ils ne permettent pas d'évaluer correctement le niveau de connaissance des apprenants. Pour remédier à ces problèmes, des études de psychologie de l'éducation ont été dédiées à la conception de QCM, et notamment sur les caractéristiques d'un QCM pertinent dans un cadre pédagogique. De ces études ont découlé différentes consignes de rédaction de QCM.

Cependant, les enseignants peuvent ne pas avoir connaissance de l'existence de ces consignes (Tarrant et Ware, 2008). Ainsi, des outils d'évaluation automatique de la qualité de QCM pourraient assister les enseignants dans leur travail de production de QCM. Pour

cette raison, nous nous intéressons à la validation automatique de QCM. Cette tâche de validation nécessite d'évaluer la qualité des différentes composantes des *items* des QCM : l'*amorce* et les *options*, ces dernières étant constituées de la *réponse* et des *distracteurs*.

Définition 1 *Un **item**, ou une **question**, est un ensemble composé d'une amorce et de ses options de réponses.*

Définition 2 *Une **amorce** (*stem*) est la consigne donnée à l'apprenant. Elle prend souvent une forme interrogative, mais peut aussi être une consigne ou un texte à trous.*

Définition 3 *Une **option** est un choix de réponse à une amorce.*

Définition 4 *Une **réponse** (*answer*) est l'option correspondant à la réponse correcte.*

Définition 5 *Un **distracteur** (*distractor*) est une option correspondant à une réponse incorrecte.*

La figure 1 présente les différentes composantes d'un item.

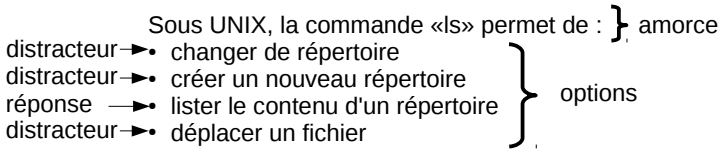


FIGURE 1 – Exemple d'item

Parmi les sous-tâches de validation des différentes composantes d'un item de QCM, celle de validation des distracteurs constitue une tâche particulièrement délicate. En effet, il s'agit d'évaluer le pouvoir discriminant des distracteurs, c'est-à-dire leur capacité à discerner les apprenants ayant compris la notion évaluée des autres apprenants. Ainsi, dans notre thèse, nous nous intéressons à la **validation automatique de distracteurs** dans le cadre de **QCM pédagogiques**. Cette tâche permettra d'assister les enseignants dans leur travail de rédaction de QCM, et peut être intégrée à la tâche de génération automatique de QCM, notamment pour la génération des distracteurs de QCM.

Les options de QCM sont de différents formats : il peut s'agir d'images, de graphiques, de nombres, mais le format d'options le plus répandu est la langue naturelle. Ainsi, dans le cadre de notre travail, nous nous intéressons uniquement aux QCM dont les options sont en langue naturelle.

PROBLÉMATIQUE ET CONTRIBUTIONS

Dans le but d'évaluer automatiquement la qualité des distracteurs de QCM pédagogiques, nous avons étudié les caractéristiques des distracteurs pertinents afin d'automatiser leur identification. L'étude de ces caractéristiques est fondée sur des consignes de rédaction de QCM, et notamment de distracteurs, développées dans le cadre de travaux de psychologie de l'éducation. Nous avons en particulier retenu la consigne suivante : **les options sont syntaxiquement et sémantiquement homogènes**, ce qui signifie que les options d'un même item sont valides si elles partagent des caractéristiques syntaxiques et sémantiques communes.

Pour évaluer la qualité des distracteurs, nous partons de l'hypothèse que les distracteurs sont rédigés en fonction de la réponse. Ainsi, estimer l'homogénéité entre les options revient à **estimer l'homogénéité entre chacun des distracteurs et la réponse**. Pour répondre à cette problématique, nous avons cherché à répondre aux trois questions suivantes :

— Comment définir l'homogénéité syntaxique et sémantique dans un but de reconnaissance automatique ?

— Quelles tâches de Traitement Automatique des Langues utiliser pour reconnaître l'homogénéité ?

— Quel modèle permet d'estimer globalement l'homogénéité ?

Pour définir l'homogénéité syntaxique et sémantique, nous nous sommes appuyé sur une analyse de QCM visant à identifier les critères d'homogénéité des options. Cela nous a conduit à proposer une définition plus complète et plus opérationnelle de l'homogénéité syntaxique et sémantique, par rapport aux travaux existants. La reconnaissance automatique de ces critères est fondée sur des analyses syntaxiques et sémantiques des options. Ces analyses permettent d'évaluer l'homogénéité syntaxique en prenant en considération des structures complexes représentant les termes comparés. Nous avons ensuite proposé d'appliquer de nouvelles mesures pour estimer l'homogénéité sémantique des options afin de tenir compte de différents types de relations sémantiques entre termes. Pour estimer globalement l'homogénéité des distracteurs, nous proposons de combiner ces mesures dans un modèle d'ordonnancement, capable de classer dans les premiers rangs les distracteurs homogènes. Nous avons proposé un cadre d'évaluation permettant de comparer nos travaux à l'état de l'art et nous montrons que nos résultats améliorent les résultats de l'état de l'art.

ORGANISATION DU MANUSCRIT

Dans le chapitre 2, nous exposons le contexte de notre travail, c'est-à-dire l'étude de QCM en psychologie de l'éducation, et nous présentons un état de l'art sur les travaux

se rapprochant du nôtre et sur lesquels nous nous fondons, soit les travaux en sélection automatique de distracteurs et différentes mesures de voisinage sémantiques dont nous utiliserons un certain nombre d'entre elles pour répondre à notre problématique.

Dans le chapitre 3, nous approfondissons la problématique de notre thèse, c'est-à-dire l'évaluation automatique de la qualité des distracteurs dans le cadre de QCM pédagogiques, et nous présentons la notion sur laquelle nous nous fondons : l'homogénéité des options. Nous définissons également le modèle que nous proposons pour répondre à notre problématique. De plus, nous présentons l'évaluation d'une analyse de corpus de QCM visant à valider la définition de l'homogénéité, et l'évaluation de méthodes de reconnaissance automatique de l'homogénéité que nous avons définies à partir de cette analyse.

Dans le chapitre 4, nous exposons l'étude des méthodes d'homogénéité sémantique que nous avons sélectionnées, et la méthode d'évaluation automatique de la qualité des distracteurs fondée sur un ordonnanceur par apprentissage. Nous présentons également les résultats de chacune de ces méthodes.

2 ÉTAT DE L'ART

Notre travail de thèse s'appuie sur les travaux effectués en psychologie de l'éducation, la branche de la psychologie dédiée à l'étude scientifique de l'apprentissage humain (Snowman, 1997). De nombreux travaux se sont notamment intéressés à l'évaluation des apprenants. Ces travaux ont notamment inspiré le développement de typologies dans le but d'évaluer les différents niveaux d'acquisition de connaissance des apprenants. Les QCM étant des outils d'évaluation des apprenants, il importe de spécifier les types d'évaluation sur lesquels nous nous focalisons. Nous présentons ces typologies à la section 2.1.

De plus, des travaux spécifiques à l'étude de QCM existent en psychologie de l'éducation. Ces travaux portent sur l'importance de rédiger des QCM de qualité, c'est-à-dire des QCM dont l'évaluation ne porte que sur l'évaluation des connaissances des apprenants, et sur les moyens d'évaluer la qualité des QCM. Afin d'aider les enseignants à rédiger ceux-ci, plusieurs ensembles de consignes de rédaction ont été développés et afin d'évaluer leur qualité, des mesures d'évaluation psychométriques ont été proposées. Nous présentons les études de QCM à la section 2.2.

Dans le but d'assister les enseignants à la rédaction de QCM, des travaux ont été dédiés à la génération automatique de QCM. Une des étapes de la génération de QCM est la sélection automatique de distracteurs, qui se rapproche de notre problématique. En effet, la tâche de sélection des distracteurs nécessite la reconnaissance de la validité des distracteurs à sélectionner, à l'instar de notre travail qui a pour objectif de reconnaître le degré de validité des distracteurs d'un item à choix multiples. La section 2.3 présente ces travaux. Pour sélectionner les distracteurs, ces travaux se fondent sur différentes mesures de voisinage sémantique. C'est pourquoi la section 2.4 présente les principales mesures de voisinage sémantique utilisées en TAL, et notamment en sélection automatique de distracteurs.

2.1 TYPOLOGIES DES QUESTIONS PÉDAGOGIQUES

Afin d'identifier les points forts et les lacunes des apprenants, des travaux en psychologie de l'éducation se sont intéressés aux différents processus cognitifs stimulés par les enseignants selon les activités qu'ils proposent aux apprenants. Ces travaux considèrent que ces processus cognitifs sont hiérarchisés selon différents niveaux d'acquisition de connaissance. Par exemple, les questions «Qu'est-ce qu'une clé primaire ?» et «Quelle est la différence entre une clé primaire et une clé étrangère ?» traitent toutes les deux du concept «clé primaire» mais la première question est une question de définition, alors

que la seconde question demande de trouver la relation entre ce concept et le concept «clé étrangère».

En psychologie de l'éducation, il existe plusieurs taxonomies représentant ces niveaux d'acquisition de connaissances. Dans ce domaine, la taxonomie la plus fréquemment citée est celle de Bloom (1956) (section 2.1.1). Elle a cependant été révisée (Krathwohl, 2002) comme nous l'expliquons dans la section 2.1.2.

Les QCM étant des outils d'évaluation, nous présentons leur couverture cognitive à la section 2.1.3 afin de positionner nos travaux selon les différents types d'évaluation.

2.1.1 *Taxonomie de Bloom*

La taxonomie la plus fréquemment citée est celle de Bloom, qui est un modèle pédagogique spécifiant différents niveaux d'apprentissage. Cette taxonomie est utilisée en génération de QCM, notamment par Cubric et Tosic (2011) qui utilisent des règles de génération catégorisées selon les niveaux de cette taxonomie. Elle est composée de six niveaux, où chaque niveau englobe tous les niveaux inférieurs. Chaque niveau correspond à des opérations spécifiques (donner une définition, appliquer un cours dans un contexte...). Selon Bloom (1956), plus l'apprenant est capable d'effectuer d'opérations, plus son niveau est élevé.

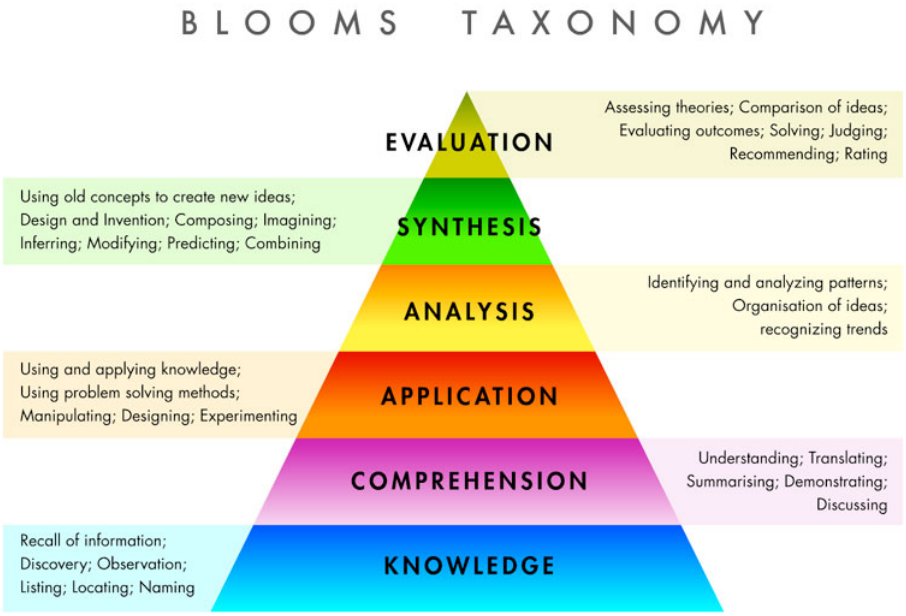


FIGURE 2 – Taxonomie de Bloom (1956)

Voici une description plus détaillée de ces niveaux, en partant du bas de la figure 2 :

1. **Connaissance** : évalue l'apprenant sur sa capacité à mémoriser son cours. La question «Quelle est la plus grande ville du Japon ?» est une question de connaissance.
2. **Compréhension** : évalue l'apprenant sur sa capacité à assimiler des concepts et à les mettre en relation avec d'autres concepts. L'apprenant est amené à effectuer des opérations telles que la démonstration, l'organisation, la comparaison ou la description de faits appris dans son cours. La question «Quelle est l'idée principale de cette histoire ?» est une question de compréhension.
3. **Application** : évalue l'apprenant sur sa capacité à appliquer les concepts du cours dans un contexte donné. La question «Comment utiliseriez-vous vos connaissances de la latitude et de la longitude pour localiser le Groenland ?» est une question d'application.
4. **Analyse** : évalue l'apprenant sur sa capacité à décomposer une information en plusieurs parties pertinentes, trouver les causes, les motivations et donner des conclusions. La question «Pourquoi les États-Unis sont-ils entrés en guerre contre l'Angleterre ?» est une question d'analyse.
5. **Synthèse** : évalue l'apprenant sur sa capacité à synthétiser plusieurs informations en les reformulant et/ou en proposant de nouvelles idées. La question «En quoi votre vie serait-elle différente si vous pouviez respirer sous l'eau ?» est une question de synthèse.
6. **Évaluation** : évalue l'apprenant sur sa capacité à présenter et défendre des opinions à partir de connaissances tout en argumentant sur ses choix. La question «Pourquoi pensez-vous que Benjamin Franklin est si célèbre ?» est une question d'évaluation.

La taxonomie de Bloom est la plus largement utilisée dans le domaine de l'éducation. Néanmoins, certains travaux en psychologie de l'éducation ont révisé cette taxonomie. La section suivante expose une des propositions de révision de la taxonomie de Bloom.

2.1.2 *Révision de la taxonomie de Bloom*

Parmi les révisions de la taxonomie de Bloom, la taxonomie de Krathwohl (2002) est l'une des plus connues. Krathwohl (2002) a développé une taxonomie bidimensionnelle fondée sur celle de Bloom. Les dimensions de sa taxonomie représentent les **connaissances** et les **processus cognitifs** attendus de la part des apprenants.

La dimension portant sur les connaissances a été construite à partir du niveau d'acquisition le plus faible de la taxonomie de Bloom, c'est-à-dire la connaissance. Cette dimension est constituée de trois grandes catégories provenant de la taxonomie originale : les connaissances factuelles, conceptuelles et procédurales. Cependant, une grande catégorie a été ajoutée : il s'agit des connaissances métacognitives, qui impliquent les connaissances sur la cognition en général.

La dimension portant sur les processus cognitifs a été construite à partir de la taxonomie de Bloom. Tout comme la taxonomie originale, elle représente les types d'opérations spécifiques pour l'évaluation des apprenants.

2.1.3 *Types d'items de QCM posés*

Dans le cadre pédagogique, les QCM sont un moyen d'évaluer les connaissances des apprenants. Les items des QCM sont associés à une connaissance et à une faculté cognitive à évaluer. Les facultés cognitives pouvant associer ces items sont le rappel de connaissances, la compréhension du domaine de connaissances à évaluer, l'application des connaissances à un contexte extérieur et l'analyse d'informations (Burton *et al.*, 1991). Ces facultés cognitives équivalent aux quatre niveaux les plus faibles de la taxonomie de Bloom (connaissance, compréhension, application et analyse) (Tarrant *et al.*, 2006; Abdalla *et al.*, 2011). Les niveaux les plus élevés de cette taxonomie, soit la synthèse et l'évaluation, ne peuvent être évalués dans des items à choix multiples car ces niveaux nécessitent de l'apprenant un travail de rédaction et d'expression d'idées personnelles, ce qui ne pourrait être formulé sous la forme d'options à sélectionner. Cependant, une étude de QCM de Tarrant *et al.* (2006) a montré que la majorité des items appellent des facultés de connaissance et de compréhension (environ 90 % de 2770 items analysés). Néanmoins, il est conseillé de rédiger des items appelant des facultés d'application et d'analyse (Burton *et al.*, 1991; Abdalla *et al.*, 2011).

Dans nos travaux, nous nous intéressons aux items de QCM associés à l'évaluation de la connaissance et de la compréhension des apprenants. En effet, il s'agit des types d'items les plus utilisés pour évaluer les apprenants.

2.2 QUESTIONNAIRES À CHOIX MULTIPLES

Les QCM sont très utilisés pour l'évaluation d'apprenants. La correction de QCM est simple et rapide, et peut être automatique : pour l'évaluateur, il s'agit de vérifier si les apprenants évalués ont sélectionné l'option correspondant à la réponse. Un grand avantage des QCM est qu'ils peuvent être utilisés pour l'auto-évaluation des apprenants. Néanmoins, si les QCM sont mal rédigés, ils peuvent provoquer un biais dans l'évaluation des apprenants : les notes obtenues peuvent ne pas refléter leur niveau réel. Pour résoudre ce problème, des consignes de rédaction de QCM ont été rédigées. À la section 2.2.1, nous présentons les méthodes d'évaluation de la qualité des QCM. À la section 2.2.2, nous présentons les consignes de rédaction de QCM proposées en psychologie de l'éducation et à la section 2.2.3, nous montrons l'influence négative des violations de ces consignes.

2.2.1 *Mesures de vérification de la qualité des QCM*

La qualité pédagogique d'un QCM, c'est-à-dire sa capacité à évaluer correctement des apprenants, est définie selon des critères psychométriques. Ces critères représentent la capacité du QCM à systématiquement évaluer les connaissances des apprenants (fiabilité, *reliability*) et sa capacité à mesurer le niveau des apprenants (validité, *validity*). Ces critères se mesurent a posteriori, c'est-à-dire qu'ils se mesurent à partir des scores obtenus par des apprenants à des tests composés d'items à choix multiples. En effet, ces critères sont fondés sur les options sélectionnées par les apprenants, ainsi que leurs notes globales. Les scores obtenus par ces mesures permettent de vérifier si un QCM est correctement rédigé ou non.

2.2.1.1 *Fiabilité*

En psychométrie, la fiabilité ou la reproductibilité d'un test est la capacité de celui-ci à mesurer systématiquement ce qu'il est supposé mesurer. Autrement dit, les résultats doivent être similaires pour des administrations répétées du même test. Dans le cadre de la vérification de la qualité des QCM, Burton *et al.* (1991) affirme que les items à choix multiples sont censés être plus fiables que les autres types de tests car les apprenants sont notés objectivement. En effet, les notes ne prennent en compte que les options sélectionnées par les apprenants, tandis que dans d'autres tests comme les questions à développement, les notes peuvent être influencées par l'écriture et la qualité de rédaction des apprenants. D'après Abdel-Hameed *et al.* (2005), la taille d'un QCM affecte sa fiabilité : plus un QCM contient d'items, plus la fiabilité est élevée. De plus, une bonne couverture du contenu à évaluer augmente la fiabilité du QCM. En revanche, un QCM dont les items se restreignent à une partie de ce contenu baisse la fiabilité car les apprenants ne sont pas notés sur l'ensemble des connaissances qu'ils sont censés avoir assimilé. La fiabilité peut être également affectée par l'état des apprenants pendant le test, ainsi que l'environnement extérieur (bruit, température...). D'après Abdel-Hameed *et al.* (2005) et Considine *et al.* (2005), la fiabilité d'un test peut être évaluée par différentes méthodes qui mesurent différents types de fiabilité : la fiabilité de test-retest (stabilité, *stability*), la fiabilité de formes équivalentes (équivalence, *equivalency*) et la cohérence interne du test (*internal-consistency*).

STABILITÉ La stabilité ou la cohérence temporelle d'un test consiste à vérifier que la variation des résultats à un test, évalué sur un même groupe de personnes, est nulle. Cela a pour objectif de vérifier qu'un test mesure bien ce qu'il est supposé mesurer, dans les mêmes conditions mais à des périodes différentes. Dans le cadre de la vérification de la rédaction de QCM, il s'agit d'évaluer plusieurs fois un même groupe d'apprenants avec le même QCM à des temps différents. D'après Abdel-Hameed *et al.* (2005) et Considine *et al.* (2005), la stabilité est mesurée en calculant un coefficient de corrélation entre les

scores des apprenants aux différentes évaluations. Cependant, le problème de cette mesure réside dans le temps d'écart entre les évaluations. Si le temps est trop court, les apprenants peuvent se souvenir des réponses sélectionnées lors de leur dernière évaluation. En revanche, si le temps est trop long, les résultats peuvent être aussi biaisés car les apprenants peuvent avoir profité de ce temps pour approfondir les connaissances du domaine évalué.

ÉQUIVALENCE L'équivalence d'un test est la comparaison de deux versions différentes d'un test (les items sont similaires sur le fond mais pas sur la forme). D'après Considine *et al.* (2005), la mesure de l'équivalence est une alternative à la mesure de la stabilité car cette dernière mesure peut être biaisée par le temps d'écart entre les différentes administrations du test. Dans le cadre de l'évaluation de la rédaction de QCM, L'objectif de la mesure d'équivalence est de vérifier si des formes alternatives de QCM mesurent de la même manière l'évaluation de mêmes connaissances. L'équivalence est mesurée en administrant les deux QCM aux apprenants. Cependant, d'après Abdel-Hameed *et al.* (2005), il peut être difficile de développer deux QCM totalement équivalents et de formes différentes.

COHÉRENCE INTERNE La cohérence interne, ou consistance interne, ou homogénéité d'un test, représente la capacité des items du test à mesurer le même concept. qui représente un domaine de connaissances particulier (par exemple, PL/SQL en bases de données). Cependant, un QCM peut couvrir plusieurs domaines de connaissances. Dans ce cas, la cohérence interne ne se mesure pas sur tout le QCM, mais seulement sur les sous-ensembles d'items portés sur le même domaine de connaissances (Miller *et al.*, 2012). Mesurer la cohérence d'interne d'un QCM (ou d'un sous-ensemble d'un QCM) revient à mesurer la corrélation entre les scores des apprenants sur les différents items du questionnaire, l'intérêt étant d'observer si les apprenants répondent de manière cohérente à différents items de l'ensemble. En effet, des apprenants maîtrisant un domaine de connaissances peuvent répondre à tous les items posés sur le même domaine de connaissances. Dans le cadre de l'évaluation de la rédaction de QCM, la mesure de la cohérence interne est la plus utilisée des mesures de la fiabilité car elle ne nécessite qu'une seule administration du test.

D'après Considine *et al.* (2005), pour qu'un QCM soit fiable, le coefficient de fiabilité doit être élevé mais pas trop car une trop forte corrélation peut montrer l'existence d'items redondants.

2.2.1.2 *Validité*

En psychométrie, la validité d'un test est le degré de celui-ci à mesurer ce qu'il est supposé mesurer. La validité est fortement liée à la fiabilité car un test valide doit être fiable. En revanche, la réciproque est fausse. Dans le cadre de l'évaluation de QCM, la

validité d'un QCM estime la capacité de celui-ci à donner des scores représentatifs du niveau de connaissances des apprenants. D'après Considine *et al.* (2005), la validité d'un QCM est généralement établie par :

— la **validité du contenu** (*content validity*), c'est-à-dire la pertinence pédagogique des items ;

— la **validité de la façade** (*face validity*), c'est-à-dire la clarté, la lisibilité et la compréhension des items pour les apprenants ;

— la **validité conceptuelle** (*construct validity*), c'est-à-dire la capacité des items à mesurer le domaine de connaissances examiné.

VALIDITÉ DU CONTENU La validité du contenu consiste à vérifier si les items d'un QCM sont appropriés et représentatifs du domaine de connaissance et des facultés cognitives évalués. Il n'existe pas de mesure calculant la validité du contenu mais celle-ci peut être estimée par des experts du domaine ayant une certaine expertise en développement de QCM.

VALIDITÉ DE LA FAÇADE La validité de la façade peut être considérée comme un sous-type de la validité du contenu. Elle porte sur la forme des items, telle que leur clarté et leur lisibilité du point de vue du contenu et de la forme (orthographe, grammaire, ponctuation et abréviations correctes). À l'instar de la validité du contenu, il n'existe pas de mesure estimant le degré de validité de la façade.

VALIDITÉ CONCEPTUELLE En psychométrie, la validité conceptuelle mesure la capacité d'un test à mesurer ce qu'il prétend mesurer. D'après Considine *et al.* (2005), la validité conceptuelle de QCM mesure la capacité d'un QCM à évaluer le niveau de connaissances du domaine de connaissance évalué. Elle est établie à partir de la **vérification de la réponse** (*key check*) et différentes analyses des options telles que l'**analyse de la difficulté de l'item** (*item difficulty analysis*), l'**analyse du pouvoir discriminant de l'item** (*item discrimination analysis*) et l'**évaluation des distracteurs** (*distractor evaluation*).

Vérification de la réponse La vérification de la réponse d'un item consiste à vérifier si la réponse est correcte et, le cas échéant, qu'il s'agit de la seule réponse correcte de l'item. Il n'existe pas de mesure psychométrique pour estimer la vérification de la réponse mais des experts du domaine examiné peuvent être habilités à vérifier la réponse.

Analyse du pouvoir discriminant de l'item L'analyse du pouvoir discriminant de l'item consiste à mesurer la manière dont les items sont liés à la performance globale du QCM. Le pouvoir discriminant d'un item part de la prémisse selon laquelle «si un item a un fort pouvoir discriminant, les scores globaux des apprenants ayant correctement répondu à cet item devraient être plus élevés que ceux des apprenants n'ayant pas correctement

répondu à cet item». Pour mesurer cette analyse, il est conseillé d'utiliser des coefficients de corrélation mesurant la relation entre le score des apprenants à l'item analysé et leurs scores globaux.

Analyse de la difficulté de l'item L'analyse de la difficulté de l'item consiste à mesurer la difficulté à laquelle les apprenants répondent à un item. Elle est mesurée par la moyenne des scores obtenus par les apprenants sur l'item. Une bonne moyenne peut montrer un bon niveau d'assimilation des connaissances ou des instructions bien comprises. En revanche, une mauvaise moyenne peut indiquer que les instructions de l'item ne sont pas appropriées ou que les apprenants n'ont pas assimilé les connaissances associées à l'item. Il est conseillé de rédiger des items de difficulté moyenne (ni trop élevée, ni trop faible).

Évaluation des distracteurs L'évaluation des distracteurs consiste à mesurer la pertinence des distracteurs d'un item. Un distracteur pertinent devrait être sélectionné par les apprenants n'ayant pas assimilé les connaissances de l'item, tandis qu'il devrait être ignoré par ceux qui ont assimilé ces connaissances. Les distracteurs qui ne sont pas ou peu sélectionnés n'ont pas d'intérêt pédagogique et devraient être supprimés ou remplacés. Si un distracteur est sélectionné plus fréquemment que la réponse, cela peut signifier que les instructions et/ou l'amorce sont mal interprétées.

Ces mesures sont conseillées pour évaluer la qualité pédagogique de QCM, comme nous le verrons dans les sections suivantes. Dans le cadre de notre travail, nous pourrons utiliser ces mesures psychométriques après avoir fait passer des QCM sur lesquels nous avons automatiquement évalué la qualité des distracteurs à des apprenants. Cependant, ces mesures étant statistiques, leur interprétation n'est pas absolue.

L'inconvénient de ces évaluations est leur non-reproductibilité : en effet, elles reposent sur des tests donnés à des apprenants, et se fondent sur les résultats des apprenants à ces tests. Ces évaluations ne peuvent donc pas s'effectuer plusieurs fois sur les mêmes apprenants avec les mêmes items. De plus, ces évaluations sont lourdes du point de vue de l'organisation et de la mobilisation d'un grand nombre d'apprenants.

Ces évaluations ont pour objectif de vérifier la qualité de QCM rédigés. Il est donc nécessaire de rédiger les items de QCM de manière à maximiser les résultats de ces évaluations. Pour assister l'enseignant dans son travail de rédaction, des travaux de psychologie de l'éducation ont été dédiés au développement d'ensembles de consignes de rédaction de QCM. Nous exposons ces travaux dans la section suivante.

2.2.2 *Consignes de rédaction de Questionnaires à Choix Multiples*

Afin d'optimiser la qualité des QCM, plusieurs ensembles de consignes de rédaction ont été proposées. Bernard et Fontaine (1982) et Burton *et al.* (1991) ont proposé une liste

d'environ 15 règles à suivre pour rédiger des QCM, Haladyna et Downing (1989) ont proposé une taxonomie de 43 règles regroupées dans des catégories liées à la construction de l'amorce, des options et à la conception du QCM. La taxonomie la plus couramment utilisée est celle de Haladyna *et al.* (2002) qui simplifie celle de Haladyna et Downing (1989) en supprimant et modifiant certaines règles, pour en arriver à 31. Celles-ci sont regroupées dans les 5 catégories suivantes : contenu de l'item (tableau 1), format de l'item (tableau 2), le style de l'item (tableau 3), rédaction de l'amorce (tableau 4) et rédaction des options (tableau 5). Dans chacun de ces tableaux, nous présentons les consignes, ainsi que des contres-exemples démontrant l'intérêt de ces consignes et la manière dont nous comptons les implémenter, dans le cas des consignes liées à la rédaction des distracteurs. L'implémentation des consignes peut être directe, c'est-à-dire qu'il peut s'agir simplement de formater les items ou prendre en considération des critères simples comme la comparaison de la longueur des options, mais peut aussi demander un processus d'élaboration complexe, c'est-à-dire un travail impliquant une problématique de recherche.

	Consigne	**Contre-exemple**	**Implémentation**
1	Chaque item doit refléter un contenu spécifique et un comportement mental unique, formulés dans des spécifications de test (grilles à deux dimensions, blueprints de tests).		Non traitée car non liée à la rédaction des distracteurs
2	Fonder chaque item sur un contenu d'apprentissage significatif ; éviter le contenu trivial.		Non traitée car non liée à la rédaction des distracteurs
3	Utiliser un matériel nouveau pour tester des apprentissages de haut niveau. Modifier en les paraphrasant les formulations des manuels ou celles utilisées par les enseignants en cours pour éviter les simples rappels de phrases par cœur.	The term *operant conditioning* refers to the learning situation in which : 1. A familiar response is associated with a new stimulus. 2. Individual associations are linked together in sequence. 3. **A response of the learner is instrumental in leading to a subsequent reinforcing event.** 4. Verbal responses are made to verbal stimuli.	Processus élaboré. Non traitée
4	Faire en sorte que chaque item ait un contenu indépendant de celui des autres.		Non traitée car non liée à la rédaction des distracteurs
5	Éviter les contenus d'items trop généraux ou trop spécifiques lors de la rédaction des items à choix multiples.		Non traitée car non liée à la rédaction des distracteurs
6	Éviter les items fondés sur l'opinion.		Non traitée car non liée à la rédaction des distracteurs
7	**Éviter les items-pièges.**		Processus élaboré. Non traitée car nécessaire de définir un item-piège
8	**Écrire les items avec un vocabulaire au niveau du public testé.**		Processus élaboré. Non traitée dans nos travaux

TABLE 1 – Consignes de rédaction de Haladyna *et al.* (2002) appartenant à la catégorie «Contenu de l'item»

	Consigne	Contre-exemple	Implémentation
9	Utiliser les versions de l'item, de la complétion et de la meilleure réponse des choix multiples conventionnels, les choix alternatifs, les vrai-faux, les multiples vrai-faux, les items dépendant d'un contexte et les ensembles d'items, mais ÉVITER les items complexes.		
10	Formater les items verticalement et non pas horizontalement.	Your supervisor informs you that three of your fifteen employees have complained to him about your inconsistent methods of supervision. The first thing you should do is 1. ask if it is proper for him to allow these employees to go over your head. 2. **ask what specific acts have been considered inconsistent.** 3. explain that you've purposely been inconsistent because of the needs of these three employees. 4. offer to attend a supervisory training program.	Implémentation directe

TABLE 2 – Consignes de rédaction de Haladyna *et al.* (2002) appartenant à la catégorie «Format de l'item»

	Consigne	Contre-exemple	Implémentation
11	Relire et tester les items avant mise en production.		
12	Écrire des phrases correctes du point de vue grammatical, orthographique, mais aussi de la ponctuation et de la typographie (capitales...).		Processus élaboré. Non traitée car nous considérons que les ressources d'entrée sont bien formées
13	**Faire en sorte de réduire le temps de lecture de chaque item.**		Processus élaboré. Non traitée car nécessaire de définir les critères de minimisation de temps de lecture et d'avoir sélectionné les distracteurs

TABLE 3 – Consignes de rédaction de Haladyna *et al.* (2002) appartenant à la catégorie «Style de l'item»

	Consigne	Contre-exemple	Implémentation
14	**S'assurer que les pistes évoquées dans l'amorce sont claires.**	California : 1. Contains the tallest mountain in the United States. 2. Has an eagle on its state flag. 3. Is the second largest state in terms of area. 4. **Was the location of the Gold Rush of 1849.**	Non traitée car non liée à la rédaction des distracteurs
15	**Inclure l'idée centrale dans l'amorce plutôt que dans les options.**		Processus élaboré. Non traitée dans nos travaux
16	**Éviter les amorces verbeuses.**	Suppose you are a mathematics professor who wants to determine whether or not your teaching of the unit on probability has had a significant effect on your students. You decide to analyze their scores from a test they took before the instruction and their scores from another exam taken after the instruction. Which of the following t-tests is appropriate to use in this situation ? 1. **Dependent samples.** 2. Heterogeneous samples. 3. Homogeneous samples. 4. Independant samples.	Non traitée car non liée à la rédaction des distracteurs
17	**Éviter les phrases comportant des négations (ne... pas..., excepté) dans les amorces. Si cela ne peut être évité, le faire avec précaution et mettre les négations en évidence (gras, capitales).**		Non traitée car non liée à la rédaction des distracteurs

TABLE 4 – Consignes de rédaction de Haladyna *et al.* (2002) appartenant à la catégorie «Rédaction de l'amorce»

	Consigne	Contre-exemple	Implémentation
18	**Écrire autant d'options pertinentes que possible, mais des recherches suggèrent que trois options sont suffisantes.**	*Obsidian* is an example of which of the following types of rocks ? 1. **Igneous.** 2. Metamorphic. 3. Sedimentary. 4. Transparent.	Implémentation directe
19	S'assurer que seulement une option correspond à la réponse.	The United States should adopt a foreign policy based on : 1. A strong army and control of the North American continent. 2. Achieving the best interest of all nations. 3. Isolation from international affairs. 4. **Naval supremacy and undisputed control of the world's sea lanes.**	Processus élaboré. Prise en compte dans nos travaux
20	**Varier l'emplacement de la réponse en fonction du nombre d'options.**		Implémentation directe
21	**Placer les options dans un ordre logique ou numérique.**		Processus élaboré. Non traitée dans nos travaux
22	Rendre les options indépendantes les unes des autres : il ne doit pas y avoir des références de l'une à l'autre.	How long does an *annual* plant generally live ? 1. **It dies after the first year.** 2. It lives for many years. 3. It lives for more than one year. 4. **It needs to be replanted each year.**	Processus élaboré. Prise en compte dans nos travaux
23	**Rendre la formulation des options homogène en contenu et en structure grammaticale.**	Idaho is widely known as : 1. **The largest producer of potatoes in the United States.** 2. The location of the tallest mountain in the United States. 3. The state with a beaver on its flag. 4. The "Treasure State."	Processus élaboré. Prise en compte dans nos travaux
24	**Rendre la longueur des phrases des options à peu près égale.**	Which of the following is the best indication of high morale in a supervisor's unit ? 1. The employees are rarely required to work overtime. 2. **The employees are willing to give first priority to attaining group objectives, subordinating any personal desires they may have.** 3. The supervisor enjoys staying late to plan the next day. 4. The unit gives expensive birthday presents to each other.	Implémentation directe
25	**L'option *aucun des choix ci-dessus* doit être utilisée avec précaution.**		Implémentation directe
26	**Éviter l'option *tous les choix ci-dessus*.**		Implémentation directe
27	Formuler les options positivement ; éviter les négations telles que *pas*.		Processus élaboré. Non traitée dans nos travaux

	Consigne	Contre-exemple	Implémentation
28	**Éviter de donner des indices menant à la réponse, tels que les suivants :** — **les adverbes tels que *toujours, jamais, complètement, absolument* ;** — **les associations par des mots qui se ressemblent ;** — **les incohérences grammaticales qui orientent vers la réponse ;** — **la réponse trop voyante ;** — **les paires ou les triplets d'options qui orientent vers la réponse ;** — **les options visiblement absurdes ou ridicules.**	To avoid infection after receiving a puncture wound to the hand, you should : 1. Always go to the immunization center to receive a tetanus shot. 2. Be treated with an antibiotic only if the wound is painful. 3. **Ensure that no foreign object has been left in the wound.** 4. Never wipe the wound with alcohol unless it is still bleeding.	Processus élaboré. Non traitée dans nos travaux
29	**Rendre plausibles tous les distracteurs.**	Which of the following artists is known for painting the ceiling of the Sistine Chapel ? 1. Warhol. 2. Flintstone. 3. **Michelangelo.** 4. Santa Claus.	Processus élaboré. Prise en compte dans nos travaux
30	Utiliser les erreurs typiques des apprenants pour écrire les distracteurs.		Processus élaboré. Non traitée car nécessite d'avoir évalué les apprenants sur les mêmes items que nous analysons
31	**Utiliser l'humour seulement s'il est compatible avec les pratiques de l'enseignant et l'environnement d'apprentissage.**		Processus élaboré. Non traitée car nécessite de définir dans les ressources pédagogiques à partir desquelles les items sont engendrés

TABLE 5 – Consignes de rédaction de Haladyna *et al.* (2002) appartenant à la catégorie «Rédaction des options»

Ces règles se rapportent à tous les points de la rédaction des items : le fond (sélection du contenu et de la faculté cognitive, homogénéité des options, inclusion de l'idée centrale de l'item dans l'amorce...) et la forme (utilisation de la grammaire et de l'orthographe correcte, ordonnancement des options selon un ordre logique ou numérique...). Néanmoins, une partie de ces règles sont générales et nécessitent des commentaires supplémentaires, ce que nous faisons au paragraphe suivant.

Dans la consigne 4 («Faire en sorte que chaque item ait un contenu indépendant de celui des autres»), l'indépendance des items signifie que la compréhension d'un item ne dépend pas de la compréhension d'autres items. Par exemple, les items dépendant de la réponse à un item précédent ne sont pas conseillés, au vu de cette règle. Dans la consigne 5 («Éviter les contenus d'items trop généraux ou trop spécifiques lors de la rédaction des items à choix multiples»), la spécificité et la généricité du contenu concerne les détails anecdotiques ou la globalité du domaine de connaissances que les apprenants doivent assimiler. La consigne 9 («Utiliser les versions de l'item, de la complétion et de la meilleure réponse des choix multiples conventionnels, les choix alternatifs, les vrai-faux, les multiples vrai-faux, les items dépendant d'un contexte et les ensembles de items, mais ÉVITER les items complexes») concerne les formats d'items à choix multiples. Ces formats sont exposés dans l'appendice A. La consigne 10 («Formater les items verticalement et non

pas horizontalement») concerne les options : le fait de les placer verticalement facilite la lecture par les apprenants car elles sont visuellement distinctes. La consigne 13 («Faire en sorte de réduire le temps de lecture pour chaque item») concerne le temps de lecture de l'amorce et des options du point de vue de l'apprenant. Elle dépend de la consigne 16 («Éviter les amorces verbeuses») car un texte long et dont une partie des informations est inutile à la compréhension de l'item peut perturber l'apprenant dans sa compréhension de l'item. La consigne 19 («S'assurer que seulement une option correspond à la réponse») implique l'importance de ne pas avoir d'options synonymes ou dont le sens de l'une est inclus dans le sens de l'autre. La consigne 20 («Varier l'emplacement de la réponse en fonction du nombre d'options») considère l'importance de l'emplacement de la réponse par rapport aux distracteurs du même item : si, dans tous les items d'un QCM, la réponse se trouve à la même position (par exemple, la 2^e option), elle peut biaiser les scores d'apprenants qui ne connaissent pas forcément leurs cours mais sélectionnent les options de même position. La consigne 24 («Rendre la longueur des phrases des options à peu près égale») fait référence au nombre de mots des options. En effet, les apprenants ne connaissant pas leurs cours auraient tendance à sélectionner l'option de taille différente des autres options du même item.

Pour élaborer notre méthode d'évaluation automatique de la qualité des distracteurs, nous nous appuyons sur ces consignes, notamment celles liées à la rédaction des options. Dans les tableaux présentant les consignes, nous avons indiqué celles que l'on peut implémenter directement et celles qui demandent une modélisation complexe. Parmi ces dernières consignes, une partie d'entre elles sont traitées dans nos travaux car elles sont directement liées à la tâche d'évaluation de la qualité des distracteurs.

La plupart des consignes de rédaction de QCM de Haladyna *et al.* (2002) sont incluses dans la taxonomie de Haladyna et Downing (1989), bien que quelques consignes de la première taxonomie soient opposées à certaines consignes de la seconde taxonomie. Ainsi, la consigne 31 de Haladyna *et al.* (2002) («Utiliser l'humour s'il est compatible avec les pratiques de l'enseignant et l'environnement d'apprentissage») est opposée à la consigne 43 de Haladyna et Downing (1989) («Éviter l'utilisation de l'humour lors du développement des options»). Les catégories de ces deux taxonomies sont différentes : celles de Haladyna et Downing (1989) séparent notamment les consignes de rédaction de la réponse et des distracteurs des consignes générales de développement des options. Le tableau 6 montre la correspondance entre les catégories de Haladyna et Downing (1989) et les consignes de Haladyna *et al.* (2002).

Nous observons que les catégories de Haladyna et Downing (1989) correspondent approximativement à celles de Haladyna *et al.* (2002), comme le montre le tableau 7.

Dans ce tableau, la catégorie «rédaction générale de l'item (procédurale)» regroupe les catégories «format de l'item» et «style de l'item». Nous constatons également que les catégories «développement des options», «développement de la réponse» et «développement

Catégories	**Consignes de Haladyna *et al.* (2002)**
Rédaction générale de l'item (procédurale)	7, 9, 10, 11, 12, 13
Rédaction générale de l'item (contenu)	1, 2, 3, 4, 5, 6, 8
Construction de l'amorce	9, 14, 15, 16, 17
Développement des options	18, 21, 22, 23, 24, 25, 26, 27, 28
Développement de la réponse	19, 20
Développement des distracteurs	29, 30, 31

TABLE 6 – Répartition des consignes de Haladyna *et al.* (2002) selon les catégories de Haladyna et Downing (1989). La consigne 18 («Écrire autant d'options pertinentes que possible, mais des recherches suggèrent que trois options sont suffisantes») correspond partiellement à la consigne 24 de Haladyna et Downing (1989) («Utiliser le plus grand nombre d'options plausibles : plus d'options sont désirables»). La consigne 31 («Utiliser l'humour s'il est compatible avec les pratiques de l'enseignant et l'environnement d'apprentissage») est opposée à la consigne 43 de Haladyna et Downing (1989) («Éviter l'utilisation de l'humour lors du développement des options»)

Catégories de Haladyna et Downing (1989)	**Catégories de Haladyna *et al.* (2002)**
Rédaction générale de l'item (procédurale)	Format de l'item, Style de l'item
Rédaction générale de l'item (contenu)	Contenu de l'item
Construction de l'amorce	Rédaction de l'amorce
Développement des options, Développement de la réponse, Développement des distracteurs	Rédaction des options

TABLE 7 – Correspondances approximatives entre les catégories de Haladyna et Downing (1989) et Haladyna *et al.* (2002)

des distracteurs» ont été regroupés en une seule catégorie : «rédaction des options».

Le but de ces consignes est de permettre une évaluation des apprenants , sans biais. Dans la section suivante, nous montrons l'impact du non-respect de ces consignes sur les scores des apprenants et nous listons les violations les plus fréquentes de ces consignes.

2.2.3 *Qualité de rédaction de Questionnaires à Choix Multiples*

Les consignes de rédaction de QCM présentées à la section précédente permettent d'engendrer des QCM évaluant des apprenants de manière optimale. Cette section montre l'impact négatif que peut avoir la rédaction de QCM de mauvaise qualité sur des apprenants. Cet impact correspond principalement à une difficulté plus ou moins grande de compréhension des items par les apprenants. Cette difficulté en plus ou en moins biaise l'évaluation : elle est donc à éviter pour noter les apprenants uniquement sur leur niveau d'assimilation du contenu du cours.

Plusieurs analyses de corpus de QCM ont montré qu'environ la moitié des items utilisés pour des examens d'apprenants comportent au moins une violation de consigne (Downing, 2005; Tarrant *et al.*, 2006).

Downing (2005) a analysé 219 items évaluant des étudiants en médecine. Parmi ces items, 100 (46 %) d'entre eux comportent au moins une violation de consigne. Les principales violations identifiées par Downing (2005) sont les suivantes :

— l'amorce n'est pas claire (violation de la consigne 15) ;

— l'amorce est de polarité négative (violation de la consigne 18) ;

— l'item comporte une option complexe (violation de la consigne 9) ;

— les options sont hétérogènes, c'est-à-dire qu'elles portent sur plusieurs domaines (violation de la consigne 24) ;

— l'option *aucune de ces réponses* est présente (violation de la consigne 26) ;

— l'option *toutes ces réponses* est présente (violation de la consigne 27).

Tarrant *et al.* (2006) a analysé 2770 items évaluant des étudiants en infirmerie. Parmi ces items, 1280 (46 %) d'entre elles comportent au moins une violation de consigne. Les principales violations identifiées par Tarrant *et al.* (2006) sont les suivantes :

— l'amorce n'est pas claire (violation de la consigne 15) ;

— l'amorce est de polarité négative (violation de la consigne 18) ;

— l'item comporte un distracteur non plausible (violation de la consigne 30) ;

— l'amorce contient des informations inintéressantes quant à sa compréhension (violation de la consigne 17) ;

— l'item contient plusieurs ou aucune réponse(s) (violation de la consigne 20) ;

— la plus longue option est la réponse (violation de la consigne 25) ;

— l'amorce contient des indices logiques sur la réponse (violation de la consigne 29) ;

— l'amorce et la réponse contiennent des mots en commun (violation de la consigne 29).

D'après ces deux analyses de QCM, les violations de consigne les plus fréquentes concernent la forme et le fond de l'amorce : dans le cas où elle est mal rédigée, son objectif n'est pas clair ou elle est de polarité négative. Les autres violations de consigne concernent la rédaction des options. Dans le cas où ils sont mal rédigés, les items comportent des indices sur la réponse, principalement du fait de l'hétérogénéité des options (les options portent sur plusieurs domaines, ou certains distracteurs ne sont pas plausibles) ou comportent plusieurs, ou aucune réponse, ce qui augmente inutilement la difficulté des items.

Concernant l'impact de ces violations de consigne sur les apprenants, des expériences ont été menées afin de le mesurer. Downing (2005) a analysé ses QCM avec les réponses d'étudiants et a classifié les items de son corpus en deux catégories : les items «standard» (sans violation de consigne) et imparfaits (comportant au moins une violation de consigne). En comparant ces deux catégories selon des critères d'évaluation de QCM (mesures psychométriques présentées à la section 2.2.1), ainsi que les notes des étudiants dans chacune de ces deux catégories, il a noté que les items imparfaits sont jusqu'à 15 % plus difficiles que les items standard, ce qui prouve que les violations de consigne rendent les items plus difficiles pour les apprenants. De plus, sur les 824 étudiants évalués, 646 (53 %) d'entre eux réussissent les examens comportant les items standard, c'est-à-dire que leur note est supérieure ou égale à un seuil, tandis que 575 (47 %) d'entre eux réussissent les examens comportant les items imparfaits. Ces 6 % d'écart montrent qu'il existe une altération (même légère) de la compréhension des items selon qu'ils comportent une violation de consigne ou non.

Ces travaux prouvent l'importance de rédiger des QCM de qualité afin d'éviter de biaiser l'évaluation des apprenants avec des erreurs de rédaction, même minimes. Dans le cadre de l'évaluation automatique de la qualité des distracteurs, il est nécessaire de prendre en compte ces consignes afin de s'assurer de la pertinence pédagogique des QCM. Les études présentées dans cette section montrent que la principale violation de consigne de rédaction de distracteurs porte sur le non-respect de l'homogénéité des options. Il est donc nécessaire d'évaluer principalement les distracteurs sur ce critère.

2.2.4 *Synthèse*

Dans cette section, nous avons présenté différentes études relatives à la conception et à l'évaluation des QCM. Nous avons montré la manière d'évaluer la qualité pédagogique d'un QCM à partir de mesures psychométriques. Nous avons présenté les différentes variantes d'items à choix multiples en montrant les variantes conseillées et déconseillées, du point de vue des enseignants et des apprenants évalués. Nous avons présenté les différentes consignes de rédaction de QCM et montré l'impact des violations de ces consignes

d'un point de vue de la qualité de l'évaluation des apprenants à travers des analyses de QCM (Downing, 2005; Tarrant *et al.*, 2006).

Dans le cadre de notre travail, nous nous intéressons particulièrement aux consignes de QCM proposées pour rédiger des QCM de qualité. En effet, nous nous appuyons sur ces consignes pour évaluer automatiquement la qualité des distracteurs, ce qui est d'autant plus important lors de la génération automatique de ceux-ci. Nous avons exposé à la section 2.2.2 les consignes que nous prenons en compte dans nos travaux. Parmi elles, nous avons défini les consignes que nous pouvons implémenter directement et celles qui demandent un processus plus complexe à modéliser. De plus, les méthodes d'évaluation pédagogique des QCM présentées à la section 2.2.1 pourront être utilisés pour évaluer nos travaux sur des ensembles d'apprenants.

Dans la section suivante, nous présentons des travaux en sélection automatique de distracteurs qui ont pour objectif de simplifier le travail de rédaction des enseignants.

2.3 SÉLECTION AUTOMATIQUE DE DISTRACTEURS

Comme expliqué à la section précédente, les QCM sont un bon moyen d'évaluer les connaissances d'apprenants. Cependant, la conception manuelle des QCM par les enseignants prend du temps et demande des efforts lors de la rédaction des différentes composantes des items à choix multiples, et notamment des distracteurs. C'est pour cette raison que des travaux se sont intéressés à la génération automatique d'items à choix multiples.

La génération automatique d'items à choix multiples est divisée en deux étapes successives : la génération de l'amorce et la sélection des distracteurs. La génération de l'amorce consiste d'abord à sélectionner la réponse à partir d'un texte source ou de ressources sémantiques, et ensuite à engendrer l'amorce. Cette étape correspond à la génération automatique de questions. La sélection des distracteurs consiste à identifier et sélectionner les éléments du texte source ou de la ressource sémantique qui pourront faire office de distracteurs de l'item. Dans cette section, nous nous intéressons aux travaux concernant la sélection automatique de distracteurs.

Les travaux que nous présentons considèrent que les distracteurs partagent des caractéristiques communes avec la réponse, notamment une homogénéité du point de vue de la forme et de la sémantique, soit suivent la consigne de rédaction n° 23 («Rendre la formulation des options homogène en contenu et en structure grammaticale»). Cette homogénéité est formalisée par une comparaison des distracteurs avec la réponse, généralement à travers différentes mesures de voisinage sémantique.

Mitkov *et al.* (2009) fondent leurs travaux sur WordNet et Wikipédia, en exploitant l'une ou l'autre ressource en fonction de la composition de la réponse (un ou plusieurs mots). Les items engendrés ont été évalués sur des cours de linguistique en langue anglaise. Si la réponse est un mot, les distracteurs potentiels sont ses coordonnées dans l'arborescence de WordNet, c'est-à-dire qu'ils partagent le même ancêtre commun direct. Si la réponse

est constituée de plusieurs mots, les distracteurs potentiels sont les syntagmes nominaux extraits des titres des pages Wikipédia dont la tête est identique à celle de la réponse. Ensuite, plusieurs stratégies de sélection ont été testées pour sélectionner automatiquement les distracteurs à partir des distracteurs potentiels, en appliquant différentes mesures de voisinage sémantique fondées sur WordNet, une mesure de voisinage distributionnel et une mesure de similarité phonétique. Les items ainsi engendrés ont été évalués par des apprenants, en calculant des mesures psychométriques. L'évaluation de ces items a montré qu'aucune de ces mesures n'est meilleure que les autres. L'une des limites de cette approche est de ne permettre la sélection de distracteurs que pour des noms et des syntagmes nominaux.

Karamanis *et al.* (2006) se sont intéressés à la génération de QCM médicaux en anglais, à partir d'un texte source et de la ressource sémantique UMLS. Les distracteurs potentiels d'une amorce sont des termes de même classe sémantique que la réponse (à partir de UMLS). Pour chacun de ces distracteurs potentiels, un score de voisinage distributionnel est calculé à partir d'un corpus de référence. Les distracteurs potentiels ayant les meilleurs scores de voisinage sémantique sont sélectionnés comme étant des distracteurs. L'évaluation des items engendrés a été effectuée par des experts du domaine qui ont jugé les distracteurs ainsi sélectionnés. Environ la moitié des items engendrés automatiquement ont été invalidés par les experts.

Papasalouros *et al.* (2008) et Cubric et Tosic (2011) fondent leurs travaux sur une ontologie de domaine. Selon la réponse sélectionnée (un concept, une instance, deux instances liées par une relation...), les distracteurs sont sélectionnés à partir de stratégies définies par des experts. Voici un exemple de stratégie :

> «Si A est un concept, a un littéral, et $A(a)$ (a appartient au concept A), alors $A(a)$ est la réponse. Pour la sélection des distracteurs, si, B est un ancêtre de A, $B(b)$, $b \neq a$, et b n'est pas un littéral appartenant à A, alors $A(b)$ est un distracteur.»

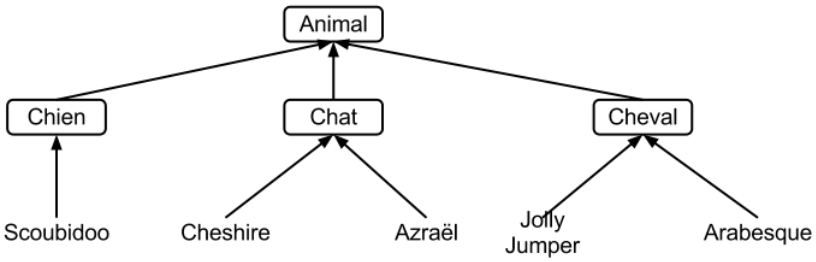


FIGURE 3 – Ontologie spécifique aux animaux célèbres

Cette stratégie permet d'engendrer l'ensemble d'options suivant, en prenant comme exemple l'ontologie de la figure 3 (A est le concept «chien» et a est le littéral «Scoubidoo») :

Amorce : Sélectionnez la phrase correcte :
Réponse : Scoubidoo est un chien
Distracteur : Azraël est un chien
Distracteur : Jolly Jumper est un chien
Distracteur : Arabesque est un chien
Distracteur : Cheshire est un chien

Cette méthode est intéressante si l'on dispose d'ontologies de domaine. Cependant, il est difficile de s'en procurer car leur conception prend du temps et requière des experts du domaine.

Lee et Seneff (2007) se sont intéressés aux QCM dont les amorces des items sont des textes à trous. Leurs travaux ont été effectués dans le cadre de l'apprentissage de la langue anglaise, et plus particulièrement pour la maîtrise des prépositions. Les distracteurs sélectionnés sont des prépositions apparaissant dans des contextes similaires à la réponse. À l'instar de Mitkov *et al.* (2009), les items engendrés ont été évaluées par des apprenants en calculant des mesures psychométriques. Ces travaux sont un peu différents des précédents car il s'agit de problèmes portant sur la langue et non sur la compréhension, et par conséquent les distracteurs sont sélectionnés en fonction de la forme des textes (existence d'une collocation) et non pas sur leur contenu sémantique.

Les principales limitations de ces travaux sont qu'ils sont limités à une application de domaine spécifique (médecine, apprentissage des prépositions) et/ou par la nature des réponses (noms et syntagmes nominaux). De plus, aucun des travaux de l'état de l'art n'évalue l'homogénéité des options sur un corpus de QCM de référence. Nous souhaitons établir une méthode générique pour l'évaluation de la qualité des distracteurs, et nous appuyer sur des corpus de QCM pour étudier et évaluer les critères que nous proposons. Afin de définir ces critères et savoir dans quelle mesure ils répondent au problème, nous proposons une définition complète de l'homogénéité des options.

Dans cette section, nous avons présenté différents travaux existants en sélection automatique de distracteurs. Ces travaux se fondent principalement sur le calcul de différentes mesures de voisinage sémantique pour sélectionner les distracteurs à partir de candidats. Dans la section suivante, nous présentons différentes mesures de voisinage sémantique, et précisons celles qui seront utiles à la résolution de notre problème.

2.4 MESURES DE VOISINAGE SÉMANTIQUE

Les mesures de voisinage sémantique quantifient le degré selon lequel des termes sont sémantiquement liés. Par exemple, les termes «palmier» et «arbre» sont sémantiquement

voisins, tout comme «palmier» et «hêtre». La reconnaissance du voisinage sémantique est nécessaire dans différentes tâches de TAL, comme la recherche d'information, la désambiguïsation lexicale, la reconnaissance de paraphrases et d'implication textuelle, la compréhension automatique de textes et, comme nous l'avons vu dans la section précédente, la sélection automatique de distracteurs, et de nombreux travaux y sont consacrés.

Les méthodes de reconnaissance de voisinage sémantique peuvent être catégorisées selon les deux principaux types de représentations sur lesquelles elles se fondent : les connaissances, sous forme de représentations sémantiques structurées, et les corpus. Les représentations structurées permettent de prendre en considération les relations sémantiques explicites pour comparer deux concepts. Cependant, les ressources correspondant à ces représentations sont souvent construites manuellement et sont souvent limitées par leur couverture. Les corpus sont des vastes ensembles de documents textuels ayant une couverture plus large et permettant de calculer des mesures de voisinage sémantique sur des bases statistiques, mais dont la nature des relations sémantiques et l'organisation hiérarchique des concepts est inconnue.

2.4.1 *Mesures fondées sur les connaissances*

Dans les représentations sémantiques structurées, les données sont organisées dans une taxonomie de concepts où les relations sont les relations sémantiques reliant ces concepts. De telles ressources représentent des connaissances spécifiques à un domaine ou des connaissances générales. Ces ressources sont des ontologies, de domaine ou non, ou des bases lexicales structurées. Dans notre travail, nous nous intéressons aux bases de connaissances générales, en domaine ouvert.

Une ontologie représentant des connaissances générales est DBpédia (Auer *et al.*, 2007). Les concepts de DBpédia représentent des entités, provenant des pages de Wikipédia, et contiennent des informations telles qu'une description (texte d'ancrage) tirée de Wikipédia, ainsi que d'autres informations telles que le type instancié par l'entité et les relations sémantiques entre les entités. Les types de DBpédia sont également des concepts de l'ontologie de DBpédia, et sont organisés entre eux par des relations hiérarchiques dans une taxonomie construite manuellement [1]. La construction des relations d'instanciation entre les entités et les types s'effectue en fonction des propriétés des objets, extraites des infoboîtes (*infoboxes*) des articles de Wikipédia associés aux entités. Les infoboîtes sont des tables préformatées de données dynamiques qui présentent sommairement des informations importantes sur une page de Wikipédia [2]. Par exemple, l'entité «Kyoto» est de type *City* car l'infoboîte de l'article de Wikipédia associé à «Kyoto» contient des informations sur la préfecture Kyoto.

1. La taxonomie de DBpédia se trouve sur la page suivante : http://mappings.dbpedia.org/server/ontology/classes/
2. http://fr.wikipedia.org/wiki/Aide:Infobox

La base lexicale structurée la plus connue est WordNet (Fellbaum, 1998), qui représente le lexique de la langue anglaise. WordNet groupe les mots synonymes en *synsets* (*synonyms sets*) liés par des relations sémantiques, hiérarchiques ou non. Une glose (définition ou description) est associée à chaque synset. Des synsets peuvent également instancier d'autres synsets, comme le synset «Paris» qui est une instance du synset «capitale nationale», mais les entités sont moins fournies que dans une ontologie comme DBpédia.

Bien que les ressources structurées que nous avons présentées soient dédiées à différents types de connaissances, ces ressources ont des structures relativement similaires : les concepts sont associés à des gloses, et sont reliés par différentes relations sémantiques, hiérarchiques ou non. Ainsi, il est possible d'appliquer les mêmes mesures de voisinage sémantique pour l'une ou l'autre de ces ressources.

Ces représentations permettent de calculer le voisinage sémantique entre deux termes (mots simples ou composés de la langue) selon les relations entre les concepts référençant ces termes (mesures fondées sur les arêtes), ou selon les informations associées aux concepts comparés (mesures fondées sur les nœuds).

2.4.1.1 *Mesures fondées sur les arêtes*

Les mesures fondées sur les arêtes estiment le voisinage sémantique entre deux concepts en calculant le plus court chemin reliant ces concepts. Elles diffèrent selon les relations retenues.

La mesure la plus classique est la distance de Rada *et al.* (1989) : elle consiste à calculer le nombre de relations hiérarchiques (hyperonymie, hyponymie, instance) du plus court chemin reliant les concepts comparés. Cette mesure a été reprise par Leacock et Chodorow (1998) (équation 1), qui tient compte de la profondeur des nœuds.

$$\mathrm{lch}(c_1, c_2) = -\log \frac{\mathrm{pcc}(c_1, c_2)}{2 \times \mathrm{MAX}} \tag{1}$$

où $\mathrm{pcc}(c_1, c_2)$ est le nombre minimal d'arêtes entre les concepts c_1 et c_2 et MAX est la profondeur de la taxonomie.

Wu et Palmer (1994) comparent la profondeur des concepts. La formule de cette mesure est la suivante :

$$wup(c_1, c_2) = \frac{2 \times \mathrm{profondeur}(\mathrm{lcs})}{\mathrm{profondeur}(c_1) + \mathrm{profondeur}(c_2)} \tag{2}$$

où $\mathrm{profondeur}(c)$ est la profondeur du concept c dans la taxonomie ($\mathrm{profondeur}(c) = 1$ si c est la racine de la taxonomie) et lcs est l'ancêtre commun le plus spécifique aux concepts c_1 et c_2. Cette mesure représente le chemin le plus court entre deux concepts selon leur profondeur, et celle de leur ancêtre commun le plus spécifique. Ce score est pondéré par la profondeur des concepts comparés dans la taxonomie. Ainsi, deux concepts profonds descendant du même parent obtiennent un meilleur score que deux concepts moins profonds de parent commun.

2.4.1.2 *Mesures fondées sur les nœuds*

Les mesures fondées sur les concepts exploitent le contenu sémantique des concepts comparés. Ce contenu peut être une valeur représentant leur importance ou une description textuelle.

Pour représenter l'importance d'un concept, un grand nombre de mesures se fondent sur la notion de *contenu informationnel*, qui calcule l'importance de l'information véhiculée par un concept dans un contexte donné. Généralement, le contexte est un corpus de documents. Ross (1976) définit le contenu informationnel, IC, d'un concept c avec la formule suivante :

$$\mathrm{IC}(c) = -\log(p(c)) \tag{3}$$

où $p(c)$ désigne la probabilité d'apparition de c dans les documents :

$$p(c) = \frac{\mathrm{freq}(c)}{N} \tag{4}$$

N désigne le nombre de mots du document apparaissant dans la ressource et la fréquence de c dans un document est :

$$\mathrm{freq}(c) = \sum_{n \in \mathrm{mots}(c)} \mathrm{occurrences}(n) \tag{5}$$

où $\mathrm{mots}(c)$ désigne l'ensemble des mots appartenant au concept c ainsi que les instances et celles des hyponymes de ce concept. Ce dernier point montre que le calcul du contenu informationnel prend aussi en considération les relations de la ressource.

Le calcul du contenu informationnel est la base de plusieurs mesures de voisinage sémantique, à l'instar des mesures de Jiang et Conrath (1997) (équation 6) et de Lin (1997) (équation 7), notamment utilisées par Mitkov *et al.* (2009) pour sélectionner des distracteurs.

$$\mathrm{jcn}(c_1, c_2) = \frac{1}{\mathrm{IC}(c_1) + \mathrm{IC}(c_2) - 2 \times \mathrm{IC}(\mathrm{lcs}(c_1, c_2))} \tag{6}$$

$$\mathrm{lin}(c_1, c_2) = \frac{2 \times \mathrm{IC}(\mathrm{lcs}(c_1, c_2))}{\mathrm{IC}(c_1) + \mathrm{IC}(c_2)} \tag{7}$$

où $\mathrm{lcs}(c_1, c_2)$ est l'hyperonyme commun le plus spécifique aux concepts c_1 et c_2.

Ces mesures comparent le contenu informationnel des concepts c_1 et c_2 avec celui de leur hyperonyme commun le plus spécifique. Ces mesures combinent les connaissances

sémantiques de ressources structurées avec les distributions des concepts dans un grand corpus. Elles privilégient les concepts dont l'hyperonyme commun le plus spécifique est proche, et dont les contenus informationnels sont proches de leur hyperonyme commun le plus spécifique.

Un autre moyen de représenter le contenu sémantique des concepts est de prendre en considération leurs descriptions textuelles. À la différence des mesures précédentes, elles ne s'appuient que sur les ressources structurées. Une mesure fondée sur la comparaison de ces descriptions est la mesure de *recoupement étendu de gloses* (Banerjee et Pedersen, 2003), notamment utilisée dans des tâches de désambiguïsation lexicale et par Mitkov *et al.* (2009) pour sélectionner des distracteurs. La mesure de recoupement étendu de gloses s'appuie sur les gloses des concepts à comparer, ainsi que celles de leurs concepts sémantiquement proches (liés par des relations directes d'hyperonymie, d'hyponymie et d'instance). Cette mesure est calculée comme suit :

$$\begin{aligned} reg(c_1, c_2) = {} & score(glose(c_1), glose(c_2)) \\ & + score(hype(c_1), hype(c_2)) \\ & + score(hypo(c_1), hypo(c_2)) \\ & + score(hype(c_1), glose(c_2)) \\ & + score(glose(c_1), hype(c_2)) \end{aligned} \tag{8}$$

où $glose(c)$ est la glose du concept c, $hype(c)$ est la glose de l'hyperonyme de c ou du concept instanciant c (si c a plusieurs hyperonymes ou est instancié par plusieurs concepts, alors $hype(c)$ est la concaténation de ces gloses) et $hypo(c)$ est la concaténation des gloses des hyponymes et des instances de c. Le score $score(g(c_1), g(c_2))$ est un score de similarité textuelle entre les gloses de c_1 et c_2 et est formulée comme suit :

$$score(g(c_1), g(c_2)) = \sum_{ch \in chaines\_communes} longueur(ch)^2 \tag{9}$$

Prenons l'exemple suivant :

> **Description de *drawing paper*** : paper that is specially prepared for use in drafting
> **Description de *decal*** : the art of transferring designs from specially prepared paper to a wood or glass or metal surface

Les chaînes communes de ces descriptions sont *paper* (longueur 1) et *specially prepared* (longueur 2). Le score est donc $1^2 + 2^2$, soit 5.

Les ressources structurées fournissent des informations sémantiques sur les concepts et sur les relations entre eux, permettant de mesurer le voisinage sémantique selon les relations entre ces concepts, et selon le contenu sémantique de ces concepts. Cependant, pour

calculer un score de voisinage sémantique entre deux concepts, cela nécessite qu'ils soient présents dans les ressources structurées, qui ont une couverture limitée du fait qu'elles se limitent à des domaines spécifiques et/ou qu'elles ont été construites manuellement. Par exemple, WordNet contient peu d'entités nommées et ne contient pas beaucoup de termes spécifiques à des domaines. Inversement, DBpédia contient peu de termes lexicaux. Il est donc nécessaire de s'appuyer sur d'autres types de représentations capables de modéliser des ressources plus vastes, même si elles ne sont pas structurées. Les corpus compensent ce manque car ils ne nécessitent pas d'être construits par des experts et qui offrent une couverture plus large que les ressources structurées.

2.4.2 *Mesures fondées sur les corpus*

Plusieurs travaux se sont fondés sur des corpus pour proposer des mesures de voisinage sémantique indépendantes de ressources structurées, et pour des langues ou des domaines où l'on dispose de vastes corpus. Contrairement aux mesures fondées sur les connaissances, qui s'appuient sur des concepts et des relations sémantiques, les mesures fondées sur les corpus emploient des approches statistiques pour représenter la sémantique des mots. Ces mesures s'appuient sur l'hypothèse distributionnelle selon laquelle des mots qui apparaissent dans des contextes similaires ont tendance à avoir des sens similaires» (Firth *et al.*, 1962).

Les mesures fondées sur les corpus s'appuient généralement sur le degré de cooccurrence des termes comparés dans les documents du corpus (section 2.4.2.1), ou sur la comparaison de leurs voisins sémantiques, qui sont des termes associés aux termes comparés dans le document (section 2.4.2.2).

2.4.2.1 *Estimation du degré de cooccurrence*

Les mesures de voisinage sémantique fondées sur le degré de cooccurrence des termes comparés s'appuient généralement sur la probabilité d'apparition de ces termes dans des documents communs. En effet, d'après l'hypothèse distributionnelle, si des termes apparaissent dans les mêmes documents, ils ont tendance à avoir des sens similaires.

Une mesure de voisinage sémantique fondée sur les cooccurrences dans les documents est PMI (*Pointwise Mutual Information*) (Turney, 2001). Cette mesure est une mesure d'association dont l'objectif est d'estimer le degré de dépendance de deux mots. La formule de PMI est la suivante :

$$\mathrm{PMI}(m_1, m_2) = \log\left(\frac{p(m_1, m_2)}{p(m_1)p(m_2)}\right) \tag{10}$$

où m_1 et m_2 sont deux mots, $p(m)$ est la probabilité d'apparition de m dans les documents, et $p(m_1, m_2)$ est la probabilité d'apparition simultanée de m_1 et m_2 dans les documents.

L'inconvénient de cette mesure est qu'elle vérifie seulement la présence des termes dans les documents. Elle ne prend pas en considération le poids (fréquence) de ces termes dans les documents. Une mesure s'appuyant sur ces poids est la mesure fondée sur l'Analyse Sémantique Explicite (*Explicit Semantic Analysis, ESA*) (Gabrilovich et Markovitch, 2007), qui calcule le voisinage sémantique entre des textes. L'ESA est une représentation vectorielle de textes dont les dimensions sont les poids du texte dans chaque document du corpus. Un mot est représenté par un vecteur de poids et un texte contenant plusieurs mots est représenté par le barycentre des vecteurs de poids représentant chaque mot du texte. Le poids d'un mot m dans un document d du corpus D correspond au TF-IDF (*term frequency-inverse document frequency*) de m dans d. Le TF-IDF (équation 11) mesure l'importance d'un terme dans un document, tout en pénalisant les termes trop fréquents dans le corpus.

$$\mathrm{tfidf}(m, d, D) = \mathrm{tf}(m, d) \times \mathrm{idf}(m, D) \tag{11}$$

où $\mathrm{tf}(m, d)$ est le nombre d'occurrences de m dans d et $\mathrm{idf}(m, D)$ (équation 12) est la fréquence inverse de m dans D.

$$\mathrm{idf}(m, D) = \log\left(\frac{|D|}{|\{d : m \in d\}|}\right) \tag{12}$$

où $|\{d : m \in d\}|$ est le nombre de documents de D où m apparaît.

La mesure de voisinage sémantique entre deux textes est une mesure de similarité vectorielle entre les vecteurs représentant les deux textes. Pour Gabrilovich et Markovitch (2007), il s'agit du cosinus des vecteurs (équation 13).

$$\cos(t_1, t_2) = \frac{\sum_i \mathrm{poids}(t_{1i})\mathrm{poids}(t_{2i})}{\sqrt{\sum_i \mathrm{poids}(t_1 i)^2 \sum_i \mathrm{poids}(t_2 i)^2}} \tag{13}$$

Une autre mesure s'appuyant sur les poids des termes est fondée sur l'Analyse Sémantique Latente (*Latent Semantic Analysis, LSA*) (Landauer et Dumais, 1997). La LSA est une représentation vectorielle des mots selon leur poids dans les documents dans une matrice, dont on réduit ensuite le rang par décomposition en valeurs singulières. Gabrilovich et Markovitch (2007) montrent que la mesure fondée sur l'ESA est plus performante pour calculer le voisinage sémantique.

2.4.2.2 *Comparaison des voisins sémantiques des termes*

Les mesures fondées sur la comparaison des voisins sémantiques s'appuient sur le fait que les termes t sont associés à d'autres termes t', dont l'ensemble de ces t' forme une représentation des t.

Les voisins sémantiques de t peuvent être les cooccurrents de t. Plusieurs mesures de voisinage sémantique sont fondées sur la comparaison des cooccurrents des termes (Dagan *et al.*, 1997; Kolb, 2008; Ferret, 2010). Nous nous intéressons à la mesure de Ferret

(2010) car il s'est intéressé au voisinage sémantique pour des applications comme la création de thésaurus distributionnels (Ferret, 2013).

Ferret (2010) extrait les cooccurrents de chaque mot m et les pondère. Pour chaque mot m du corpus, ses cooccurrents sont extraits et pondérés selon l'intensité de leur lien avec m. Ces poids sont regroupés sous la forme de vecteurs représentant m. Ces vecteurs permettent de calculer le voisinage sémantique entre deux mots. Chaque mot m du corpus est représenté sous la forme d'un vecteur de cooccurrents obtenu en comptabilisant les cooccurrences observées entre m et les mots d'une fenêtre de taille fixe centrée sur toutes les occurrences de m dans le corpus. Le poids de chaque cooccurrent c est calculé par une mesure d'association entre m et c. L'évaluation de sa méthode montre que l'information mutuelle (équation 10) est la mesure la plus performante. La mesure de voisinage sémantique de deux mots, soit deux vecteurs de cooccurrents, est une mesure de similarité vectorielle. L'évaluation de la méthode de Ferret (2010) a montré que le cosinus est la mesure la plus performante (cf. équation 13).

Les voisins sémantiques de t peuvent également être des termes-clés présents dans la description de t. Cependant, cela nécessite des ressources dont les documents sont des descriptions de termes, comme Wikipédia. À partir de cette ressource, il est possible d'identifier les termes-clés des descriptions par l'ensemble des liens entrants et sortants associés aux pages Wikipédia. Cette ressource est structurée mais n'indique pas la nature des relations sémantiques entre les différentes pages. Une mesure de voisinage sémantique fondée sur de telles représentations a été proposée par Milne et Witten (2013) : l'outil qu'ils ont développé, Wikipedia Miner[3], calcule un score appris sur ces liens à partir de dumps de Wikipédia. Ce score est une combinaison de huit attributs représentant quatre mesures calculées sur les liens entrants et sortants :

— l'union des liens des pages comparées ;

— l'intersection des liens des pages comparées ;

— la *normalized link distance*, adaptée de la *normalized Google distance* (Cilibrasi et Vitanyi, 2007), calculant la distance sémantique de deux pages p_1 et p_2 en comparant les pages de Wikipédia où apparaissent les liens associés à p_1 et p_2. Si p_1 et p_2 sont liées aux mêmes pages, cela signifie un fort voisinage sémantique entre p_1 et p_2, tandis que si p_1 et p_2 sont liées à des pages différentes, cela signifie un faible voisinage sémantique entre p_1 et p_2. La formule de la *normalized link distance* est la suivante :

$$\mathrm{nld}(p_1, p_2) = \frac{\log(\max(|P_1|, |P_2|) - \log(|P_1 \cap P_2|))}{\log(|W|) - \log(\min(|P_1|, |P_2|))} \tag{14}$$

où P_1 et P_2 sont les ensembles de pages reliant respectivement p_1 et p_2, et W est l'ensemble de toutes les pages de Wikipédia ;

3. http://wikipedia-miner.cms.waikato.ac.nz/

— la similarité vectorielle des liens (*link vector similarity*), inspirée de la mesure du TF-IDF mais appliquée aux liens des pages comparées vers les pages de Wikipédia, au lieu des mots traités par le TF-IDF. Les dimensions des vecteurs comparés sont calculées à partir du *lf* × *iaf* (*link frequency* × *inverse article frequency*). La fréquence des liens (*link frequency*) donne une estimation de l'importance du lien l_i dans une page p_j :

$$lf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{15}$$

où $n_{i,j}$ est le nombre d'occurrences du lien l_i dans la page p_j. La fréquence inverse d'article (*inverse article frequency*) mesure l'importance générale d'un lien :

$$iaf_i = \log(\frac{|W|}{|p : l_i \in p|}) \tag{16}$$

où W est l'ensemble des pages de Wikipédia. Le score de voisinage sémantique est l'angle des vecteurs des pages comparées.

La mesure de Milne et Witten (2013) permet de calculer le score de voisinage sémantique de termes s'il existe des pages Wikipédia associées. Cette mesure a l'avantage de ne prendre en considération que les termes importants, annotés par des utilisateurs, représentés par les liens entrants et sortants des pages Wikipédia. De plus, cette ressource permet de reconnaître des termes synonymes, c'est-à-dire des termes référés par la même page Wikipédia.

Les mesures de voisinage sémantique fondées sur les représentations contextuelles permettent de comparer la sémantique de termes à partir de vastes corpus de documents non annotés, selon des méthodes statistiques. Ces mesures permettent aussi bien de reconnaître des synonymes que des termes de sens proche.

2.4.3 *Synthèse*

Dans cette section, nous avons présenté différents types de mesures de voisinage sémantique : les mesures fondées sur des représentations structurées, et les mesures fondées sur des représentations contextuelles. Les mesures fondées sur des représentations structurées permettent de reconnaître efficacement le voisinage sémantique de deux concepts dont on connaît le sens, mais la couverture des ressources structurées reste limitée. En revanche, les mesures fondées sur des représentations contextuelles permet de comparer de mots ou de documents sans restriction lexicale, mais la nature des relations liant ces mots ou documents est inconnue.

Pour répondre à notre problématique, il nous semble important de prendre en considération des représentations structurées pour mesurer efficacement le voisinage sémantique

entre options, mais nous ne souhaitons pas nous limiter à un domaine spécifique. Ainsi, nous prendrons également en considération des représentations contextuelles pour mesurer le voisinage sémantique de tout type d'option.

2.5 SYNTHÈSE

Dans ce chapitre, nous avons observé que les QCM correspondent aux quatre niveaux d'acquisition de connaissances les plus faibles de la taxonomie de Bloom. Cependant, pour l'évaluation de la qualité des distracteurs, les questions d'application et d'analyse nécessitent des ressources supplémentaires en plus de documents référençant les items. Ainsi, nous ne nous intéresserons qu'aux questions de connaissance et de compréhension. Nous avons indiqué les consignes de rédaction de QCM qui orienteront nos travaux de recherche à la section 2.2.2 : «*Rendre les options indépendantes les unes des autres : il ne doit pas y avoir des références de l'une à l'autre*» et «*Rendre la formulation des options homogène en contenu et en structure grammaticale*». Les mesures psychométriques d'évaluation de QCM pourront être utilisées pour évaluer ceux qui seront engendrés à partir de nos travaux afin de mesurer leur qualité pédagogique.

Par rapport aux quelques travaux existants en sélection automatique de distracteurs, nos travaux se rapprochent de ceux de Karamanis *et al.* (2006) et Mitkov *et al.* (2009) car nous souhaitons prendre en compte un texte source et des ressources sémantiques pour sélectionner les distracteurs, ainsi que différentes mesures de voisinage sémantique pour comparer les distracteurs. Cependant, nous ne nous limitons pas à un domaine : nous nous intéressons à l'évaluation de la qualité des distracteurs indépendamment d'un domaine. De ce fait, nous ne nous limitons pas à une seule ressource sémantique. Nous nous appuyons à la fois sur des ressources structurées et distributionnelles à partir desquelles nous calculons les mesures de voisinage sémantique.

Dans le chapitre suivant, nous présentons notre problématique et l'approche que nous proposons à partir des travaux présentés dans le chapitre présent.

3

HOMOGÉNÉITÉ DES DISTRACTEURS : DÉFINITION ET MODÉLISATION

Les QCM sont couramment utilisés dans différents contextes d'apprentissage et d'évaluation d'apprenants humains car ils sont des indicateurs objectifs du niveau de connaissance des apprenants. De plus, il est possible de les corriger automatiquement. Cependant, la rédaction de QCM prend un temps considérable, et la qualité des QCM est cruciale afin de s'assurer que les apprenants seront évalués uniquement en fonction de leurs connaissances. Dans le chapitre 2, nous avons présenté des consignes de rédaction de QCM développées pour aider à la création de QCM de qualité. Néanmoins, les QCM rédigés peuvent présenter des défauts car les enseignants peuvent ne pas avoir connaissance de ces consignes, ou peuvent avoir des difficultés pour les appliquer. Ainsi, une évaluation automatique de la qualité de QCM pourrait aider les enseignants dans leur tâche de rédaction.

Lors de la rédaction de QCM, la sélection des distracteurs est une tâche difficile. C'est pourquoi nous sommes intéressé à établir la qualité des distracteurs. Nous proposons donc une méthode d'*évaluation automatique de la qualité des distracteurs*.

Nous souhaitons vérifier s'il est possible de prendre en compte les consignes de rédaction de distracteurs (cf. section 2.2) pour évaluer automatiquement leur qualité. Pour cela, nous avons appuyé notre travail sur l'étude d'un corpus de QCM.

Avant de présenter notre approche, nous allons rappeler les éléments suivants de l'item à choix multiples : une *amorce*, la *réponse* à cette amorce et les *distracteurs* à évaluer. Étant donné que l'on traite des QCM d'évaluation de connaissance et de compréhension de texte, ces items sont rédigés à partir d'un document de référence ou attendent une réponse se trouvant dans un texte contenant les notions évaluées. Ainsi, nous prenons également en compte le *document de référence* de l'item pour obtenir des informations supplémentaires sur les distracteurs et la réponse, s'ils sont présents dans le document. L'exemple 1 illustre les éléments présentés dans ce paragraphe.

Document : In Africa, nine out of ten people with HIV/AIDS are still denied these drugs, now almost universally available in wealthy countries. The reason ? Lack of political will and *high drug prices*. Universal access to treatment is an achievable goal, but it requires the United States and EU to act at this month's World Trade Organization meeting to respect poor countries' right to import cheaper generic versions of AIDS drugs.
Amorce : What is the economic reason for the almost lack of access to ARV drugs for patients in Africa ?
Réponse : high drug prices
Distracteur : availability of ARVs in wealthy countries
Distracteur : external debt cancellation for the African governments
Distracteur : profits of pharmaceutical companies
Distracteur : lack of political plans

Exemple 1 – Item à choix multiples

Notre objectif est d'estimer automatiquement le degré de validité des distracteurs. Pour ce faire, nous nous sommes fondé sur les consignes de rédaction de QCM développées en psychologie de l'éducation, et notamment les consignes 19 et 23 présentées dans le chapitre 2 : «*S'assurer que seulement une option correspond à la réponse*» et «*Rendre la formulation des options homogène en contenu et en structure grammaticale*». Ces consignes indiquent que des options valides sont des options dont les structures syntaxiques et les sens des options sont assez proches, mais dont les sens ne sont pas tout à fait similaires. Nous proposons donc une **définition** de l'*homogénéité* des options pour évaluer automatiquement la qualité des distracteurs.

Définition 6 *L'**homogénéité** correspond aux caractéristiques communes entre un distracteur et les autres options, faisant en sorte que ce distracteur pourrait correspondre à une réponse possible à l'amorce, mais aussi à ce qui différencie le distracteur des autres options. Cette notion d'homogénéité peut être déclinée selon deux aspects : l'aspect syntaxique et l'aspect sémantique.*

La pertinence d'un distracteur est évaluée en comparant certaines caractéristiques de celui-ci avec celles des autres options de l'item. Cependant, un distracteur est pertinent s'il correspond (partiellement) aux informations attendues par l'amorce. Cette correspondance peut être établie en comparant la réponse et l'amorce. La pertinence d'un distracteur est donc principalement fondée sur son homogénéité avec la réponse. L'homogénéité entre un distracteur et les autres options est donc ramenée pour simplification à l'homogénéité entre le distracteur et la réponse. Pour la même raison, nous ne considérons pas l'amorce pour estimer la pertinence des distracteurs.

Par ailleurs, notre objectif est de définir un modèle générique d'évaluation de la qualité des distracteurs. Notre travail ne sera donc pas dépendant d'un domaine spécifique ou d'autres caractéristiques comme le niveau des apprenants évalués.

Nous ne traitons que des items dont les réponses sont courtes car ce type de réponse permet d'exploiter des ressources sémantiques. Ces réponses sont des entités nommées, des chunks (cf. définition 7) et des syntagmes nominaux. Les réponses longues amèneraient à proposer des méthodes différentes.

Par ailleurs, l'analyse des QCM que nous avons effectuée a montré que ce type d'option est très fréquent (cf. section 3.4).

Définition 7 *Un **chunk** est la plus petite séquence d'unité linguistique possible formant un groupe avec une tête forte, et qui n'est ni discontinue, ni récursive (Abney, 1992).*

Nous considérons des chunks étendus :

— **syntagme nominal** : syntagme dont le noyau est un nom («le petit chien» et «le médecin de Jeanne» sont des syntagmes nominaux). Nous nous restreignons à l'étude des syntagmes nominaux constitués :
  — d'un chunk nominal («le petit chien») ;
  — d'un chunk nominal suivi d'un chunk prépositionnel («le médecin de Jeanne») ;
  — d'un chunk nominal, suivi d'un chunk prépositionnel, suivi d'un chunk nominal («les profits des entreprises pharmaceutiques»).

— **syntagme prépositionnel** : syntagme constitué d'une préposition et d'un *chunk* nominal («par le gouvernement» est un syntagme prépositionnel) ;

— **chunk verbal** : chunk constitué d'un verbe («pris», «partant» et «a appris» sont des chunks verbaux) ;

— **chunk adjectival** : chunk constitué d'un adjectif et de ses modifieurs éventuels et n'appartenant pas à un syntagme nominal («vrai» et «plus facile» sont des chunks adjectivaux) ;

— **chunk adverbial** : chunk constitué d'un adverbe n'appartenant pas à un autre chunk ou un syntagme nominal («prochainement» et «le plus» sont des chunks adverbiaux).

Les entités nommées sont des syntagmes nominaux référant à un nom de personne, de lieu ou d'organisation. Étant donné qu'elles possèdent des informations sémantiques supplémentaires, elles seront traitées différemment des autres syntagmes nominaux. Les types de chunks que nous traitons étendent les types de termes traités dans l'état de l'art : en effet, ces travaux se restreignent aux noms communs et aux syntagmes nominaux.

Dans ce chapitre, nous présentons le modèle que nous proposons pour évaluer automatiquement la qualité des distracteurs (section 3.1), puis nous précisons la définition de l'homogénéité que nous avons présentée en introduction de ce chapitre (section 3.2). Nous présentons ensuite les corpus sur lesquels nous nous sommes appuyé pour valider la définition de l'homogénéité et évaluer le modèle (section 3.3). Nous validons cette définition par une analyse de corpus et nous évaluons des méthodes de reconnaissance de l'homogénéité syntaxique et sémantique (section 3.4).

3.1 MODÈLE

Le modèle que nous proposons pour évaluer la qualité des *distracteurs* est un modèle supervisé d'ordonnancement, appris automatiquement, capable d'ordonner un ensemble de *candidats*, dont une partie correspond aux distracteurs à évaluer, et dont l'autre partie sont des termes de même nature syntaxique que la réponse. Notre objectif est de reconnaître le degré d'homogénéité des distracteurs par rapport à la réponse pour un item. Le degré d'homogénéité d'un distracteur est reconnu selon son classement parmi l'ensemble des candidats. Un distracteur est d'autant plus pertinent s'il est classé en tête parmi les candidats (cf. figure 4).

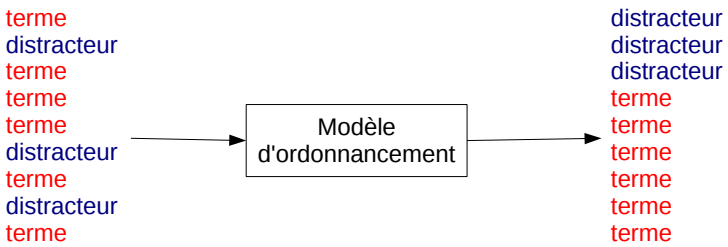


FIGURE 4 – Modèle d'ordonnancement de candidats

Selon les cas d'utilisation, les termes peuvent provenir du document de référence ou d'une banque de distracteurs disponible, par exemple.

Pour répondre à notre problématique, un modèle d'ordonnancement est plus pertinent qu'un modèle de classification binaire, où l'on catégorise les distracteurs comme étant valides ou non. La raison en est que selon la nature de la réponse, l'homogénéité des distracteurs peut être plus ou moins forte. Des distracteurs peuvent être pertinents même si leur homogénéité avec la réponse n'est pas élevée. Un distracteur est pertinent si son homogénéité avec la réponse est plus importante que l'homogénéité entre les autres candidats et la réponse. L'ordonnanceur calcule un score d'homogénéité $s(c, r)$ entre chacun des candidats c et la réponse r. Ce score est calculé selon différentes mesures d'homogénéité sémantique, présentées au chapitre 4.

Les candidats sont sélectionnés selon des critères d'homogénéité syntaxique. Nous avons choisi de dissocier l'évaluation de l'homogénéité syntaxique et sémantique car les critères sont très différents.

Notre travail est le premier à proposer un modèle d'ordonnancement appris et évalué sur des corpus de QCM et à permettre de combiner différentes mesures d'homogénéité sémantique. En effet, nous avons vu au chapitre 2 que les méthodes proposées ne se

fondent que sur des mesures de voisinage sémantique individuelles, et évaluent leurs méthodes sur des scores d'apprenants.

3.2 HOMOGÉNÉITÉ DES DISTRACTEURS

Comme nous l'avons expliqué dans l'introduction de ce chapitre, l'évaluation de la qualité des distracteurs dépend de l'homogénéité entre les distracteurs et la réponse. Cette notion d'homogénéité se décline selon l'aspect syntaxique (section 3.2.1) et sémantique (section 3.2.2). Dans la suite du document, lorsque nous voulons faire référence à la réalisation en langue naturelle des candidats ou de la réponse, nous la nommerons *terme*, c'est-à-dire un mot ou un groupe de mots.

3.2.1 *Homogénéité syntaxique*

Pour qu'un distracteur soit valide d'un point de vue syntaxique, sa structure syntaxique doit correspondre aux informations attendues par l'amorce et doit être proche de celle de la réponse. Il est donc possible d'estimer le degré de pertinence syntaxique d'un candidat à travers son homogénéité syntaxique avec la réponse sans vérifier directement la correspondance entre la structure syntaxique du candidat et l'amorce.

L'homogénéité syntaxique signifie que les distracteurs ont (au moins en partie) une structure syntaxique proche de celle de la réponse.

Amorce : De quel pays est originaire le kimchi ?
Réponse : Corée
Distracteur : Japon
Distracteur : Chine
Distracteur : Mongolie

Exemple 2 – Item à choix multiples

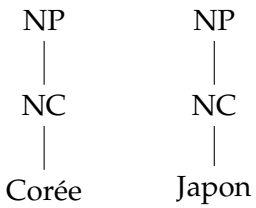


FIGURE 5 – Arbres de constituants des options de l'exemple 2

Amorce : De quel type de blessure a souffert le premier ministre Letton ?
Réponse : Une blessure non mortelle
Distracteur : Une blessure potentiellement mortelle
Distracteur : Une égratignure

Exemple 3 – Item à choix multiples

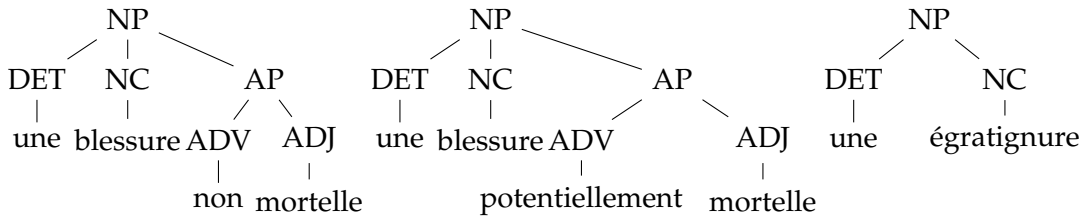


FIGURE 6 – Arbres de constituants des options de l'exemple 3

Une représentation de la forme syntaxique d'un énoncé est l'arbre de constituants. Cet arbre représente les différents éléments syntaxiques de la phrase organisés de manière hiérarchique. Par exemple, les options des exemples 2 et 3 (figures 5 et 6) sont toutes constituées de syntagmes nominaux, comme le montrent leurs analyses syntaxiques. Elles sont donc syntaxiquement homogènes.

Nous nous fondons sur ce type d'arbre syntaxique pour mesurer l'homogénéité syntaxique.

Pour évaluer l'homogénéité syntaxique entre un candidat et une réponse, nous prenons en considération plusieurs critères relatifs aux arbres de constituants des options comparées.

Un des critères d'homogénéité syntaxique porte sur la comparaison des structures globales du candidat et de la réponse. Dans un arbre de constituants, la structure globale est le nœud racine, que l'on appelera *type syntaxique* du candidat ou de la réponse dans la suite du manuscrit. Dans les figures 5 et 6, toutes les options sont des syntagmes nominaux (NP). Ce critère consiste à vérifier si les types syntaxiques du candidat et de la réponse sont similaires – soit identiques ou assez proches pour que le candidat puisse correspondre à une réponse possible de l'amorce – et donc de déclasser les candidats non homogènes, à l'instar du distracteur de l'exemple 4. En effet, les distracteurs ont généralement le même type syntaxique que la réponse car ce type correspond à l'information syntaxique principale attendue par l'amorce (cf. figure 7). Ce critère est donc binaire : les types syntaxiques du candidat et de la réponse sont similaires ou non.

Amorce : De quel pays est originaire le kimchi ?
Réponse : Corée
Distracteur : Voyager au Japon

Exemple 4 – Item à choix multiples dont le distracteur est syntaxiquement non valide

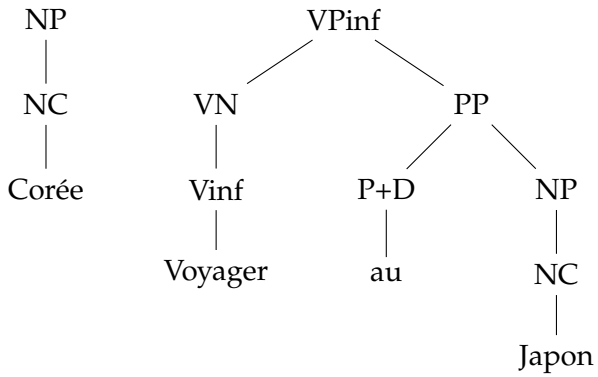


FIGURE 7 – Arbres de constituants des options de l'exemple 4

Le critère que nous avons présenté est intéressant pour filtrer les candidats non homogènes. Ce filtrage permet d'éviter le calcul inutile de scores d'homogénéité à de tels candidats.

Cependant, d'autres critères d'homogénéité syntaxique sont nécessaires pour calculer le degré d'homogénéité des candidats restants. Ces critères se fondent sur les arbres syntaxiques des candidats et de la réponse. L'exemple 5 montre l'intérêt de s'appuyer sur d'autres critères syntaxiques.

Amorce : De quel pays est originaire le kimchi ?
Réponse : Corée
Disracteur : Un pays où il fait bon vivre

Exemple 5 – Item à choix multiples

Bien que ce distracteur et cette réponse soient des syntagmes nominaux, ils sont faiblement homogènes car leurs représentations syntaxiques sont considérablement différentes (cf. figure 8).

Dans notre modèle, nous nous appuyons sur la notion d'homogénéité syntaxique pour filtrer les candidats non homogènes, c'est-à-dire des candidats de type syntaxique différent de celui de la réponse. Nous proposons une définition plus complète de l'homogénéité syntaxique que les travaux existants en sélection de distracteurs. En effet, ceux-ci ne traitent que d'options de même type syntaxique, donc des options syntaxiquement homogènes par nature.

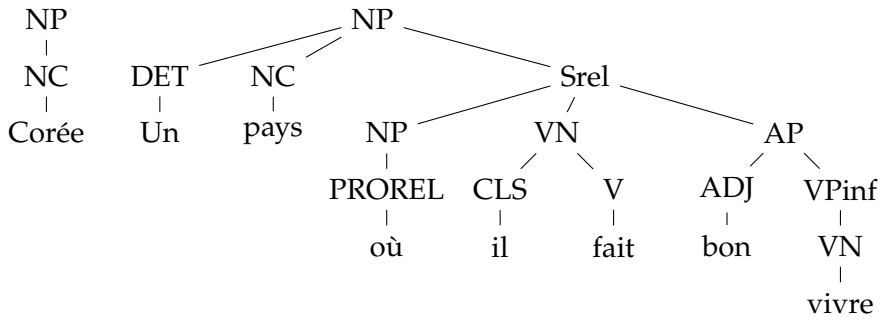


FIGURE 8 – Arbres de constituants des options de l'exemple 5

3.2.2 *Homogénéité sémantique*

Un candidat et une réponse sont sémantiquement homogènes s'ils ont suffisamment de caractéristiques sémantiques communes. Ainsi, dans l'exemple 2, toutes les options correspondent à des noms de pays asiatiques. Dans l'exemple 3, toutes les options correspondent à des types de blessures. Plus généralement, deux termes sont sémantiquement homogènes si les concepts auxquels ils se réfèrent peuvent être englobés par un concept général.

Pour définir la notion d'homogénéité sémantique, nous ferons référence à une organisation hiérarchique des connaissances. Nous avons choisi de nous référer à l'ontologie WordNet (que nous avons présenté à la section 2.4), dont quelques extraits sont présentés aux figures 9 et 10. WordNet est constitué de concepts et de relations sémantiques, dont nous présentons celles que nous exploitons dans la liste ci-dessous.

— **synonymie** : deux mots m_1 et m_2 sont synonymes s'ils font référence au même concept c («Biélorussie» est un synonyme de «Bélarus») ;

— **instanciation** : un concept c_1 est une instance d'un concept c_2 si c_2 correspond au type sémantique de c_1 («France» est une instance de «pays européen») ;

— **hyperonymie** : un concept c_1 est un hyperonyme d'un concept c_2 si c_1 a un sens plus général que c_2 («lieu» est un hyperonyme de «pays») ;

— **hyponymie** : l'inverse de l'hyperonymie ;

— **holonymie** : un concept c_1 est un holonyme d'un concept c_2 si c_1 désigne un ensemble comprenant c_2 («Europe» est un holonyme de «France») ;

— **méronymie** : l'inverse de l'holonymie.

Dans la suite du manuscrit, nous ne ferons référence aux relations que dans un sens (le sens ascendant) : nous conserverons donc les relations d'instanciation, d'hyponymie et de méronymie.

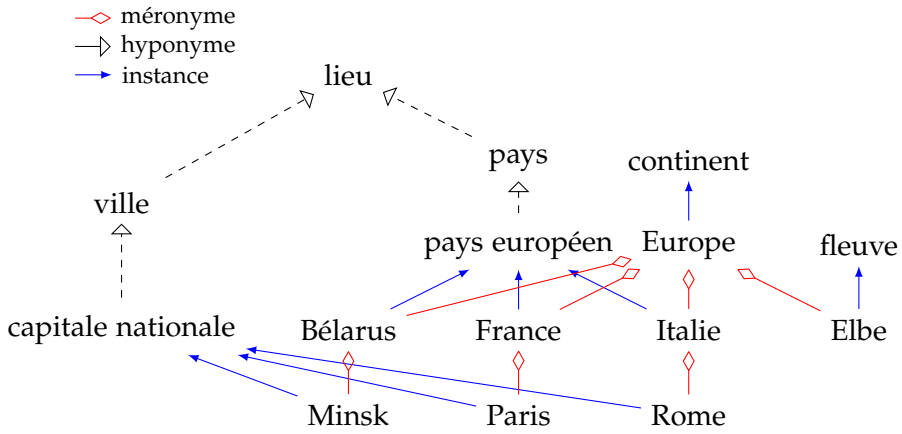


FIGURE 9 – Caractérisation sémantique de paires de nœuds de type entité nommée. Ce graphe est la traduction en français d'un extrait de la ressource WordNet

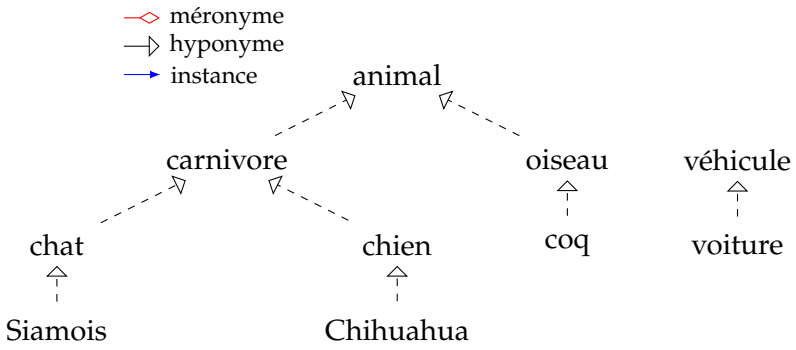


FIGURE 10 – Caractérisation sémantique de paires de nœuds de type chunk non entité nommée. Ce graphe est la traduction en français d'un extrait de la ressource WordNet

Nous donnons maintenant la définition des notions nécessaires à la définition de l'homogénéité sémantique entre une réponse et un candidat, donc entre deux termes, en faisant référence à WordNet pour les exemples. Ces notions sont le *voisinage sémantique*, la *similarité sémantique* et la *spécificité sémantique*.

Définition 8 *Le **voisinage sémantique** indique dans quelle mesure deux concepts sont sémantiquement distants dans un réseau sémantique en utilisant toutes les relations entre eux.* [1]

Deux termes sont sémantiquement voisins lorsqu'il existe un chemin entre les concepts auxquels ils se réfèrent, et le degré de voisinage peut être vérifié en fonction de la longueur du chemin. Dans la figure 9, le terme «France» est un voisin sémantique de tous les autres concepts : le chemin entre «France» et les autres concepts ne dépasse pas trois. Dans la figure 10, le concept «chien» est un voisin sémantique de tous les autres concepts, excepté «voiture» et «véhicule» car il n'existe pas de lien sémantique entre «chien» et ces concepts.

Définition 9 *La **similarité sémantique** est un cas particulier de voisinage sémantique : deux termes sont similaires s'ils ont le même sens ou si le sens d'un des termes est inclus dans le sens du second terme, c'est-à-dire que les concepts auxquels ils se réfèrent sont liés par une chaîne ascendante ou descente de relations d'instanciation, d'hyponymie ou de méronymie.*

Généralement, la reconnaissance de la similarité sémantique ne prend pas en considération les relations de méronymie. Seules les relations d'instanciation et d'hyponymie sont considérées pour cette tâche. Nous y avons ajouté les relations de méronymie afin de nous conformer à la consigne de QCM numéro 19 : «*S'assurer que seulement une option correspond à la réponse*».

Dans la figure 9, les concepts «France» et «Europe» sont similaires car «France» est un méronyme de «Europe». Dans la figure 10, les concepts «chien» et «carnivore» sont similaires car «chien» est un hyponyme de «carnivore».

La notion d'*homogénéité sémantique* est proche de la notion de *voisinage sémantique* puisque deux termes sont sémantiquement homogènes s'ils partagent un grand nombre de propriétés sémantiques communes.

Cependant, le voisinage sémantique englobe la similarité sémantique. Nous avons vu que des distracteurs similaires à la réponse ne sont pas recommandés car ils correspondent (au moins partiellement) à des réponses correctes et peuvent donc rendre un test plus difficile pour les apprenants. En prenant comme exemple la figure 11 et l'exemple 6, de potentiels distracteurs comme «pays européen» et «Europe» ne seraient pas valides car ces deux derniers concepts sont similaires à «France» (cf. figure 13).

1. La définition que nous donnons est reprise de la définition du voisinage sémantique par Ponzetto et Strube (2007) : "Semantic relatedness indicates how much two concepts are semantically distant in a network or taxonomy by using all relations between them (i.e. hyponymic/hypernymic, antonymic, meronymic and any kind of functional relations including *is-made-of*, *is-an-attribute-of*, etc.)"

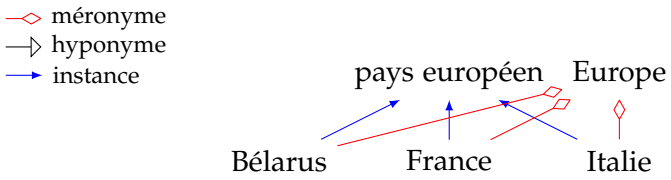


FIGURE 11 – Caractérisation sémantique de paires de nœuds

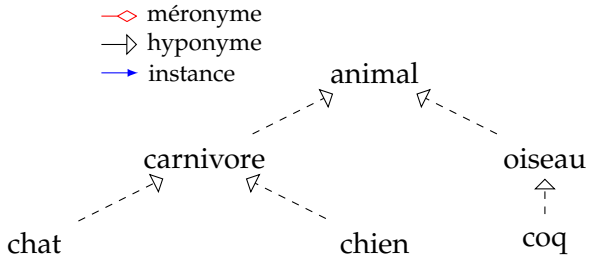


FIGURE 12 – Caractérisation sémantique de paires de nœuds

Amorce : Où se trouve le Musée du Louvre ?
Réponse : France

Exemple 6 – Exemple de couple d'amorce-réponse

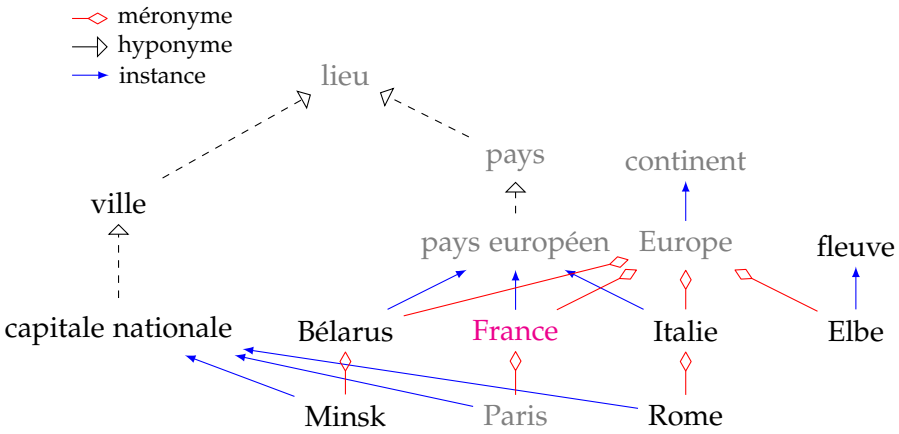


FIGURE 13 – Caractérisation sémantique de paires de nœuds. Les nœuds gris représentent les concepts similaires au concept «France»

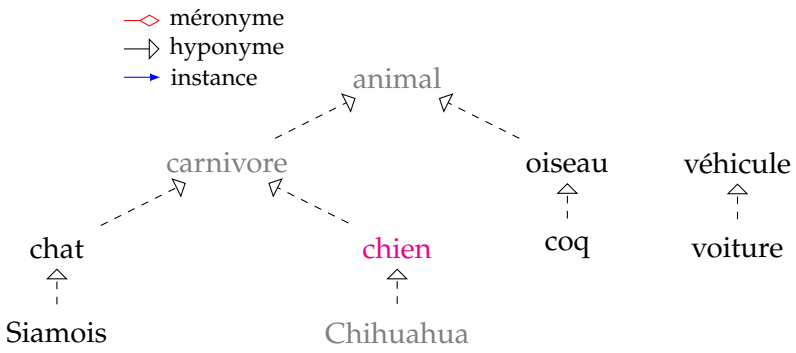


FIGURE 14 – Caractérisation sémantique de paires de nœuds. Les nœuds gris représentent les concepts similaires au concept «chien»

Pour prendre un autre exemple où les options ne sont pas des entités nommées, comme la figure 12 et l'exemple 7, de potentiels distracteurs comme «carnivore» et «Chihuahua» ne seraient pas valides car ces deux derniers concepts sont similaires à «chien» (cf. figure 14).

Amorce : Quel être vivant ronge des os ?
Réponse : chien

Exemple 7 – Exemple de couple d'amorce-réponse

De ce fait, la notion de similarité sémantique doit être exclue de celle de l'homogénéité sémantique.

Cependant, l'exclusion de la similarité sémantique du voisinage sémantique ne suffit pas à définir l'homogénéité sémantique. En effet, des candidats peuvent être voisins et non similaires à la réponse, et ne pas être plausibles. En prenant comme exemple la figure 13 et l'exemple 6, un distracteur comme «Elbe» n'est pas plausible, bien qu'il soit un voisin de la réponse «France» et non similaire à cette réponse. La raison est que ce distracteur ne partage pas d'ancêtre proche de la réponse. Nous introduisons donc la notion de *spécificité sémantique*, qui correspond à la distance entre les concepts et leur ancêtre commun. La spécificité est maximale si l'ancêtre commun des concepts comparés est direct, tandis qu'elle est faible si l'ancêtre commun de ces concepts est éloigné, et inexistante si ces concepts ne partageant pas d'ancêtre commun.

En prenant comme exemple la figure 15 et l'exemple 6, les candidats les plus plausibles sont «Bélarus» et «Italie» car ils sont des pays, et des candidats comme «Minsk» et «Rome» le sont moins car ils ne sont pas des pays. En revanche, «Elbe» n'est pas un candidat plausible car il ne partage pas d'ancêtre commun direct avec «France».

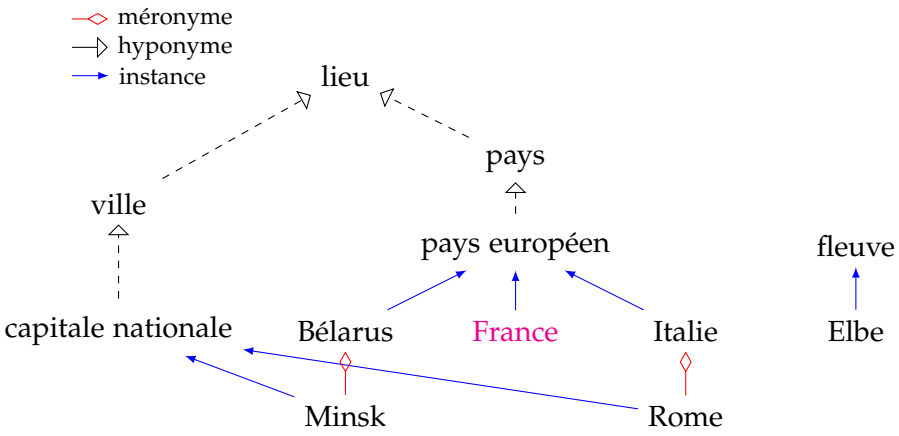


FIGURE 15 – Caractérisation sémantique de paires de nœuds

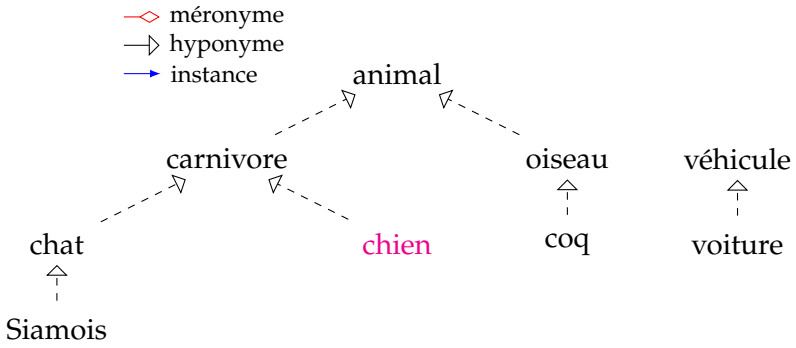


FIGURE 16 – Caractérisation sémantique de paires de nœuds

Amorce : Quel animal ronge des os ?
Réponse : chien

Exemple 8 – Exemple de couple d'amorce-réponse

Pour prendre un exemple où les options ne sont pas des entités nommées, comme la figure 16 et l'exemple 8, le candidat le plus plausible est «chat», «coq» est un candidat moins plausible car il s'agit d'un animal non carnivore et «voiture» n'est pas plausible car il ne s'agit pas d'un animal.

Les distracteurs doivent ainsi correspondre au type attendu par l'amorce. Étant donné que les informations sémantiques de la réponse correspondent ainsi à ce type attendu, la correspondance entre les candidats et le type attendu par l'amorce peut s'établir par le degré de spécificité sémantique entre les distracteurs et la réponse.

La notion de spécificité sémantique est importante pour estimer le degré de validité des candidats mais ne permet pas de les invalider.

À partir des notions que nous avons présentées dans cette section, nous pouvons maintenant définir l'homogénéité sémantique.

Définition 10 *L'**homogénéité sémantique** est un cas particulier de voisinage sémantique qui exclut la notion de similarité sémantique. Enfin, une meilleure homogénéité est atteinte si le degré de spécificité sémantique est élevé.*

Notre définition de l'homogénéité sémantique s'appuie sur une organisation structurée des connaissances. Cependant, comme nous l'avons expliqué dans la section 2.4, de telles ressources ne sont pas toujours disponibles et elles ont une couverture limitée. Pour estimer l'homogénéité sémantique, il est nécessaire de s'adapter à d'autres ressources existantes. Dans la section 2.4, nous avons vu que l'utilisation de corpus est un bon moyen de compenser le manque de couverture des ressources structurées pour estimer le voisinage sémantique entre deux termes. Nous proposons donc de combiner l'utilisation de ressources structurées et de corpus pour estimer automatiquement l'homogénéité sémantique par des mesures que nous détaillons au chapitre 4. Nous introduisons de nouvelles mesures basées sur celles proposées par Mitkov *et al.* (2009), ainsi qu'une combinaison de mesures fondées sur différents types de ressources afin d'étendre la couverture et mesurer des propriétés de natures différentes. Les travaux existant en sélection automatique de distracteurs (présentés en section 2.3) ne s'appuient que sur une seule mesure pour estimer l'homogénéité, et sont donc limités par les inconvénients de l'utilisation de l'une ou l'autre ressource.

En ce qui concerne la similarité entre deux termes, nous avons vu qu'elle peut être identifiée à partir des relations entre les concepts référencés par ces termes, dans des ressources structurées. Cependant, il est nécessaire de reconnaître la similarité entre les

termes non présents dans de telles ressources. Les mesures de voisinage sémantique fondées sur les corpus (cf. section 2.4) ne permettent pas d'identifier les relations hiérarchiques mais peuvent donner une indication sur les relations de synonymie : un score de voisinage sémantique élevé, voire maximal, peut donner une indication sur la similarité de deux termes.

3.3 CORPUS RECUEILLIS

Pour estimer automatiquement la qualité des distracteurs, nous nous appuyons sur des corpus de QCM provenant de différentes sources.

Nous avons principalement travaillé sur un corpus de QCM en langue anglaise. Dans l'optique de faire varier les types de QCM, nous avons collecté ceux-ci à partir de différentes sources :

— des évaluations de systèmes de compréhension automatique de textes fournis par QA4MRE (Peñas *et al.*, 2013) [2]. Notre corpus contient des questions des campagnes QA4MRE 2011, 2012 et 2013 (ensemble qa4mre) ;

— plusieurs sites d'apprentissage de la langue anglaise. Ces QCM ont pour objectif d'évaluer des apprenants sur la compréhension de la langue (ensemble evalAnglais) ou sur l'évaluation de connaissances (ensemble evalConnaissances).

Les deux ensembles du corpus sont relativement homogènes entre eux : les QCM de l'ensemble qa4mre sont destinés à l'évaluation de systèmes de compréhension automatique de textes et ceux de l'ensemble evalAnglais est destiné à des apprenants adultes de la langue anglaise.

Préalablement à la proposition du modèle que nous avons présenté dans le chapitre précédent, nous avons effectué plusieurs analyses de distracteurs dans le but de valider l'homogénéité entre distracteurs et réponses et préciser sa définition. Pour cela, nous avons travaillé sur un échantillon du corpus (corpus qcmValidation, voir le tableau 8 pour les informations détaillées).

Pour apprendre et évaluer notre modèle, nous nous sommes appuyé sur un autre échantillon de QCM (qcmModele, voir le tableau 8 pour les informations détaillées).

À partir du corpus qcmModele, nous avons extrait deux sous-ensembles constitués des items dont la réponse est une entité nommée (corpus qcmEN, voir le tableau 8 pour les informations détaillées) et des items dont la réponse est un chunk non entité nommée (corpus qcmNonEN, voir le tableau 8 pour les informations détaillées). Les items du corpus qcmEN ont été sélectionnés manuellement. Les items du corpus qcmNonEN ont été sélectionnés automatiquement, à partir d'une analyse syntaxique effectuée par l'outil Stanford Parser (Klein et Manning, 2003), en appliquant les critères d'identification de chunk exposés en début de chapitre.

2. QA4MRE (Question Answering for Machine Reading) est une tâche proposée par la campagne d'évaluation CLEF (Conference and Labs of the Evaluation Forum), http://www.clef-initiative.eu/

corpus	ensemble	# it.	# op.	(# op.)/it.	prop. # it. / qcmEN	objectif
qcmValidation	qa4mre	100	500	5		compréhension automatique de textes
	evalAnglais	68	298	4,4		évaluation de la langue
	evalConnaissances	20	80	4		évaluation de connaissances
	total	188	878	4,7		
qcmModele	qa4mre	341	1531	4,5		compréhension automatique de textes
	evalAnglais	394	1292	3,3		évaluation de la langue
	total	735	2823	3,8		
qcmEN	qa4mre	56	252	4,5		compréhension automatique de textes
	evalAnglais	47	150	3,2		évaluation de la langue
	total	103	402	3,9	14,0 %	
qcmNonEN	qa4mre	74	328	4,4		compréhension automatique de textes
	evalAnglais	122	413	3,4		évaluation de la langue
	total	196	741	3,8	26,7 %	

TABLE 8 – Caractéristiques des corpus en langue anglaise

Le corpus qcmNonEN contient des options des différents types de chunks que nous avons présenté en introduction de ce chapitre. Le tableau 9 montre la répartition des options de ce corpus selon leurs types de chunk. Nous observons que la grande majorité des options de ce corpus sont des syntagmes nominaux.

	Nombre	**Pourcentage**
syntagmes nominaux	538	72,6 %
chunks adjectivaux	71	9,6 %
chunks verbaux	65	8,8 %
chunks adverbiaux	34	4,6 %
chunks prépositionnels	33	4,5 %

TABLE 9 – Répartition des options du corpus qcmNonEN selon leur type de chunk

Les corpus qcmValidation et qcmModele ont été construits à partir de sources communes. Par conséquent, une partie des items du corpus qcmValidation sont également présents dans le corpus qcmModele. Les deux corpus possèdent 23 items en commun, soit 12 % du corpus qcmValidation et 3 % du corpus qcmModele. La figure 17 montre les relations entre les différents corpus présentés.

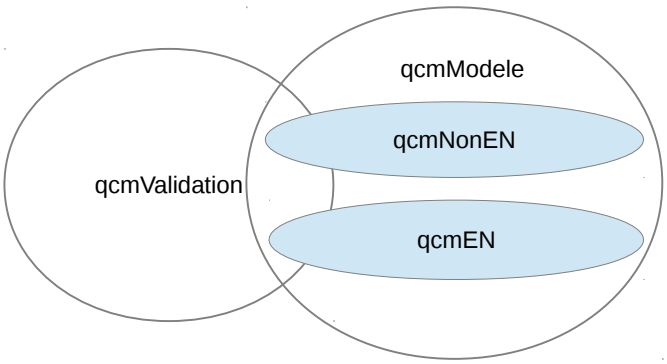


FIGURE 17 – Relations entre les différents corpus de QCM

L'analyse de corpus a principalement porté sur l'analyse de l'homogénéité syntaxique. De plus, les mesures d'homogénéité sémantique du modèle sont différentes de celles étudiées lors de l'analyse de corpus. Par conséquent, le fait d'avoir des items similaires dans les corpus qcmModele et qcmValidation ne provoque pas de biais dans l'apprentissage.

Étant donné la faible taille du corpus qcmModele, nous ne l'avons pas séparé en un corpus d'apprentissage et de test, mais appliquons un processus de validation croisée pour évaluer notre méthode.

Dans l'objectif de vérifier la validité de notre modèle dans d'autres langues, nous avons également constitué un corpus de QCM en langue française (qcmModeleFr, voir le tableau 10 pour les informations détaillées). Cependant, ce corpus est moins hétérogène que le corpus de QCM en langue anglaise. En effet, l'intégralité des QCM de ce corpus sont des QCM de compréhension de la langue française. Nous avons constitué un sous-corpus dont les réponses des items sont des chunks non entité nommée (qcmNonENFr, voir le tableau 10 pour les informations détaillées). Étant donné qu'il existe moins de ressources sémantiques structurées en français qu'en anglais, nous n'avons pas pu élaborer un modèle complet et avons effectué quelques expériences préliminaires (cf. chapitre 5) en français.

corpus	**ensemble**	**# it.**	**# op.**	**(# op.)/it.**	**objectif**
qcmModeleFr	evalFrançais	336	1020	2,8	évaluation de la langue
qcmNonENFr	evalFrançais	220	552	2,5	évaluation de la langue

TABLE 10 – Caractéristiques des corpus en langue française

3.4 VALIDATION DE L'HOMOGÉNÉITÉ

Pour valider la possibilité d'évaluer automatiquement l'homogénéité des distracteurs, nous avons effectué une analyse de QCM. Pour ce faire, nous sommes parti d'un corpus de QCM où les distracteurs ont été annotés manuellement en fonction de leur degré d'homogénéité avec la réponse, selon des critères syntaxiques et sémantiques. Nous avons automatisé cette annotation afin de vérifier la possibilité de reconnaître automatiquement l'homogénéité des distracteurs. Les méthodes d'annotation automatique que nous proposons s'appuient sur les analyses syntaxiques des options analysées, ainsi que sur des annotations sémantiques générales.

L'évaluation de ces annotations automatiques a montré qu'il est possible d'estimer l'homogénéité des distracteurs. Cette évaluation nous a conduit à développer le modèle d'évaluation de la qualité des distracteurs (cf. section 3.1). Nous présenterons le modèle d'évaluation de distracteurs au chapitre 4.

Dans les sections suivantes, nous présentons chacune des annotations en donnant leurs résultats afin de pouvoir les analyser. Les annotations sont fondées sur :

— la similarité des arbres syntaxiques (section 3.4.2) ;

— la conformité de l'option au type attendu par l'amorce (section 3.4.3) ;
— la similarité des types d'entité nommée (section 3.4.4).

Les différentes analyses d'annotation des distracteurs suivent une méthodologie d'évaluation similaire que nous présentons au préalable (section 3.4.1).

3.4.1 *Méthodologie d'évaluation*

Les annotations manuelles ont été conçues et effectuées lors d'un stage (André, 2013), sur le corpus qcmValidation. Nous avons ensuite évalué des méthodes de reconnaissance automatique de l'homogénéité des distracteurs en annotant automatiquement ce même corpus, afin de vérifier si les annotations automatiques correspondaient bien aux annotations manuelles.

Ces méthodes s'appuient sur une analyse syntaxique et une annotation en entités nommées des options. L'analyse syntaxique s'est effectuée avec l'outil Stanford Parseret l'annotation en entités nommées avec l'outil Stanford Named Entity Recognition (Finkel *et al.*, 2005).

Pour chacune des méthodes évaluées, nous avons comparé les annotations automatiques aux annotations manuelles en calculant le rappel (équation 17), la précision (équation 18), et la f-mesure (équation 19) pour chacune des catégories d'annotation cat. Ces catégories sont présentées dans les sections décrivant les critères de validation de la définition de l'homogénéité (sections 3.4.2, 3.4.3 et 3.4.4).

$$R(cat) = \frac{\#\text{ distracteurs correctement catégorisés}}{\#\text{ distracteurs catégorisés automatiquement dans } cat} \tag{17}$$

$$P(cat) = \frac{\#\text{ distracteurs correctement catégorisés}}{\#\text{ distracteurs catégorisés manuellement dans } cat} \tag{18}$$

$$F(cat) = 2\frac{R(cat) \times P(cat)}{P(cat) + P(cat)} \tag{19}$$

Cette analyse de corpus a donné lieu à une publication (Pho *et al.*, 2014).

3.4.2 *Similarité des structures syntaxiques*

Dans la section 3.2.1, nous avons expliqué que l'homogénéité syntaxique des termes se traduit par la similarité de leurs structures syntaxiques. Nous avons voulu vérifier que cette caractéristique se retrouve dans les QCM. Pour cela, nous sommes parti d'un corpus annoté manuellement, selon le degré de similarité entre les distracteurs et la réponse, indépendamment du sens des options. Nous l'avons ensuite annoté automatiquement pour vérifier qu'il est possible de reconnaître automatiquement l'homogénéité syntaxique.

3.4.2.1 *Annotation manuelle des distracteurs*

Les réponses peuvent être exprimées en différentes structures syntaxiques, souvent relatives à la forme de l'amorce : elles peuvent être des entités nommées, des syntagmes (nominaux ou verbaux), des clauses ou des phrases. Ainsi, des critères syntaxiques ont été définis sur la base des chunks, pour permettre de comparer ces différentes structures syntaxiques. Les distracteurs sont classifiés selon quatre catégories :

— **syntaxe identique :** représente les distracteurs constitués des séquences de chunks identiques à celle de la réponse. Par exemple, le distracteur «The number of tortoises began to decrease»

«**NP**(The number) **PP**(of tortoises) **VP**(began) **VP**(to decrease)»

and the answer «The number of tortoises began to grow»

«**NP**(The number) **PP**(of tortoises) **VP**(begin) **VP**(to grow)»

ont une syntaxe identique : leur séquence de chunks est «*NP PP VP VP*» ;

— **syntaxe partiellement identique :** représente les distracteurs qui partagent la même séquence de chunks que leurs réponses associées, mais avec une variation (insertion, suppression ou substitution de chunk). Par exemple, le distracteur «it resists diseases»

«**NP**(it) **VP**(resists) **NP**(diseases)»

et la réponse «it is not profitable»

«**NP**(it) **VP**(is not) **ADJP**(profitable)»

ont une variation : le dernier ADJP de la réponse est remplacé par un NP dans le distracteur ;

— **syntaxe globalement identique :** représente les distracteurs ayant plus d'une variation de chunks par rapport à la réponse, mais partageant la même structure globale que la réponse, c'est-à-dire le même type de clauses ou de syntagmes de plus haut niveau. Par exemple, le distracteur «because the amount of CO_2 saved by using renewable energies is not considered» et la réponse «because they only consider current emissions but not previous ones» diffèrent de plus d'une variation, mais ces options sont toutes les deux des clauses causales subordonnées : la syntaxe est considérée comme étant globalement identique ;

— **syntaxe différente :** représente les distracteurs ne partageant pas la même syntaxe globale que la réponse. Par exemple, le distracteur «military operations and migrant labor» et la réponse «leveraging financial funds and financing HIV/AIDS programs for Africa» ont une syntaxe différente : le distracteur est une coordination de syntagmes nominaux tandis que la réponse est une coordination de syntagmes verbaux.

3.4.2.2 *Annotation automatique des distracteurs*

Afin de classifier automatiquement les distracteurs selon ces catégories, nous avons comparé les arbres de constituants des réponses et des distracteurs, à différents niveaux de généralité : celui de l'étiquetage morpho-syntaxique, des chunks, des nœuds de haut niveau des arbres de constituants, et de l'ensemble de l'arbre. Pour classifier les distracteurs, nous avons calculé la distance de Levenshtein entre sa séquence de chunks et celle de la réponse.

- Si la distance vaut 0, nous considérons que le distracteur et la réponse ont une syntaxe identique.
- Si la distance vaut 1 (coût d'une opération), nous considérons que le distracteur et la réponse ont une syntaxe partiellement identique.
- Si la distance est supérieure à 1, nous comparons les nœuds de plus haut niveau des arbres de constituants du distracteur et de la réponse. Si ceux-ci partagent la même séquence de nœuds à ce niveau, nous considérons qu'ils ont une syntaxe globalement identique.
- Si aucune des conditions présentées ci-dessus n'est respectée, le distracteur et la réponse ont une syntaxe différente.

3.4.2.3 *Évaluation des annotations*

Afin d'évaluer l'homogénéité syntaxique des distracteurs, 490 distracteurs ont été annotés, sur les 650 distracteurs du corpus qcmValidation. Les exemples redondants ont été éliminés.

ANNOTATION MANUELLE Le tableau 11 montre la répartition des distracteurs selon les différentes catégories syntaxiques.

	Nombre	**Pourcentage**
Syntaxe identique	189	39,5 %
Syntaxe partiellement identique	91	19,0 %
Syntaxe globalement identique	141	29,4 %
Syntaxe différente	58	12,1 %
Total	490	100,0 %

TABLE 11 – Répartition des distracteurs selon leur homogénéité syntaxique avec la réponse

D'après ces résultats, nous observons qu'environ 40 % des distracteurs annotés partagent une syntaxe commune avec la réponse. Ces distracteurs sont principalement des

entités nommées mais quelques phrases et clauses appartiennent à cette catégorie. Les distracteurs restants sont principalement le résultat d'une substitution du verbe ou du sujet par rapport à la structure syntaxique de la réponse.

La moitié des distracteurs reconnus comme ayant une «syntaxe partiellement identique» à celles de la réponse sont des listes ou des distracteurs dont la réponse est une liste, comme l'exemple 9.

Réponse : Concurrent List, Union List, Residuary Subject List
Distracteur : Union and State List

Exemple 9 – Couple de réponse-distracteur

De plus, une partie des distracteurs appartenant à cette catégorie peuvent effectivement présenter de légères variations syntaxiques avec la réponse, mais restent assez similaires.

Quasiment tous les distracteurs reconnus comme ayant une «syntaxe globalement identique» à celle de la réponse sont des clauses. La raison est que leurs structures syntaxiques ne suivent pas une homogénéité syntaxique aussi stricte que les distracteurs comme les chunks ou les entités nommées. La dernière catégorie, «syntaxe différente», ne contient pas beaucoup de distracteurs (12 %).

Ces résultats valident l'existence de l'homogénéité syntaxique entre distracteurs et réponses. 40 % d'entre eux ont une syntaxe identique à celle de la réponse, tandis qu'environ 90 % des distracteurs ont une syntaxe proche de celle de la réponse si l'on relâche la contrainte d'identité stricte.

ANNOTATION AUTOMATIQUE Le tableau 12 montre les résultats de l'évaluation de l'annotation automatique des distracteurs selon le critère syntaxique.

	Précision	**Rappel**	**F-mesure**
Syntaxe identique	0,71	0,83	0,77
Syntaxe partiellement identique	0,50	0,30	0,38
Syntaxe globalement identique	0,58	0,67	0,62
Syntaxe différente	0,28	0,36	0,31

TABLE 12 – Résultats de l'évaluation de l'annotation syntaxique automatique

Nous observons que les distracteurs de syntaxe identique à la réponse sont largement reconnus par l'annotation automatique, et plus de la moitié des distracteurs de syntaxe globalement identique à la réponse sont reconnus. Cependant, les distracteurs de syntaxe partiellement identique ou différente de la réponse ne sont pas très bien reconnus. Une partie de ces distracteurs sont manuellement annotés comme étant de «syntaxe globalement identique» et annotés automatiquement comme étant de «syntaxe partiellement

identique», ou vice-versa. Une partie moins importante des distracteurs annotés manuellement comme étant de «syntaxe partiellement identique» sont annotés automatiquement comme étant de «syntaxe identique» car leurs séquences de chunks identifiées (à partir des arbres syntaxiques) sont similaires, comme le montre l'exemple 10.

Réponse : burning forests

«**VP**(burning) **NP**(forests)»

Distracteur : growing new species

«**VP**(growing) **NP**(new species)»

Exemple 10 – Couple de réponse-distracteur

Plus de la moitié des distracteurs de syntaxe différente ont été reconnus comme étant de «syntaxe partiellement différente». Cela est dû au fait que l'annotation automatique ne prend en considération que les chunks et les nœuds racine, et pas les structures syntaxiques dans leur totalité.

Réponse : spontaneous fires

«**NP**(spontaneous fires)»

Distracteur : reducing forests

«**VP**(reducing) **NP**(forests)»

Exemple 11 – Couple de réponse-distracteur

Dans l'exemple 11, nous observons que les structures syntaxiques sont différentes, mais la distance de Levenshtein sur les séquences de chunks est de 1.

Pour vérifier si les tailles des distracteurs influencent l'homogénéité syntaxique, nous avons également comparé les séquences des chunks et calculé la distance d'édition sur les arbres représentant les distracteurs et la réponse (Zhang et Shasha, 1989). La figure 18 montre la distance de Levenshtein entre les chunks des distracteurs et des réponses en fonction des longueurs des distracteurs en nombre de mots.

Cette figure montre qu'une grande partie des distracteurs n'ont pas une longueur élevée (moins de 10 mots) et que les séquences de chunks des distracteurs de longueur très faible (entre 1 et 3 mots) varient très peu en fonction de leurs réponses associées. Afin d'analyser les autres distracteurs, nous avons pris en compte le nombre moyen de mots par chunk (environ 2 mots). Nous observons généralement une variation partielle des séquences de chunks des distracteurs composés de plus de 3 mots par rapport à leurs

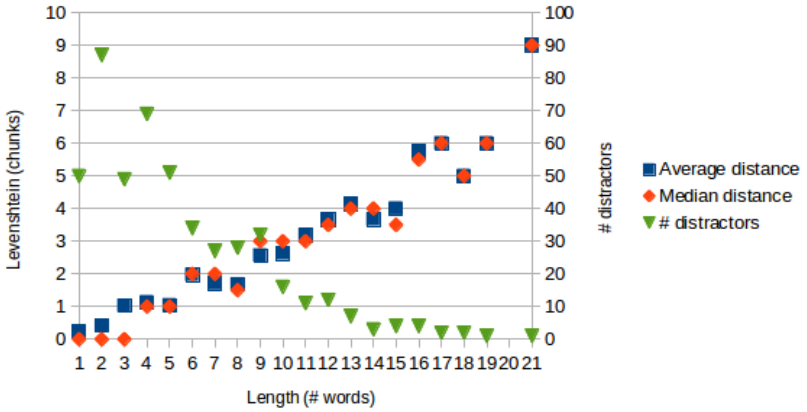


FIGURE 18 – Distance de Levenshtein entre les chunks des distracteurs et de leurs réponses associées en fonction des longueurs des distracteurs, et nombre de distracteurs par longueur

réponses associées (en fonction du nombre moyen de mots par chunks, entre un et deux tiers des mots). D'après le tableau 11, ces distracteurs ont une syntaxe partiellement ou globalement identique à celle de la réponse.

Afin de comparer les structures syntaxiques, nous avons supprimé les mots de ces arbres, représentés par les feuilles. Nous avons calculé la distance d'édition sur les arbres syntaxiques (sans les mots) des distracteurs et de la réponse. Cette distance calcule le coût minimal des opérations (insertions, suppressions et substitutions de nœuds) pour transformer un arbre syntaxique en un autre. Comparativement à l'analyse précédente qui a permis d'observer les variations des niveaux inférieurs des structures syntaxiques, l'analyse de la distance d'édition sur les arbres permet d'observer les variations sur la totalité des arbres de constituants. La figure 19 montre la distance d'édition sur les arbres entre les arbres de constituants des distracteurs et de leurs réponses associées en fonction des longueurs de ces distracteurs.

Nous observons qu'une partie des structures syntaxiques des distracteurs est partagée avec celle de la réponse. De plus, à l'instar des distances de Levenshtein entre les distracteurs et les réponses, la distance d'édition sur les arbres augmente fortement dans le cas des longs distracteurs (plus de 10 mots). Cela montre que l'homogénéité syntaxique est respectée pour créer des distracteurs, spécialement dans le cas des distracteurs courts tels que les entités nommées et les chunks.

L'analyse de corpus a permis de valider la définition de l'homogénéité syntaxique sur tout type d'option. Nous avons également montré qu'il est possible de reconnaître automatiquement l'homogénéité syntaxique en comparant les structures syntaxiques des op-

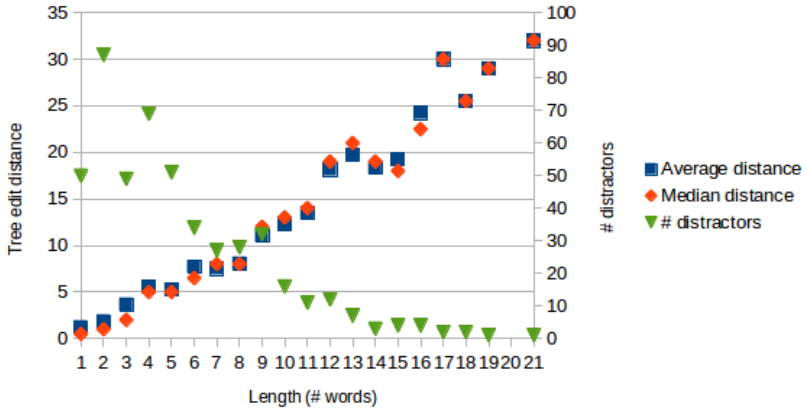


FIGURE 19 – Distance d'édition sur les arbres entre les arbres de constituants des distracteurs et de leurs réponses associées, et nombre de distracteurs par longueur

tions. Dans le modèle, nous prenons en considération le critère d'homogénéité syntaxique pour invalider les candidats de syntaxe différente de la réponse.

Les analyses suivantes ont pour but de valider la définition de l'homogénéité sémantique. À l'instar de l'analyse de l'homogénéité syntaxique, nous analysons tout type d'option.

3.4.3 *Conformité de l'option au type attendu par l'amorce*

L'homogénéité sémantique des options peut être estimée selon leur conformité sémantique avec le type attendu par l'amorce. Ce type peut être un *type spécifique*, indiqué explicitement dans l'amorce (par exemple, «Which president had the most children ?» attend d'une réponse qu'elle soit une personne qui est un «président») ; ou un *type d'entité nommée* (classiquement un nom de personne, de lieu ou d'organisation. Par exemple, «Who invented the telephone ?» attend une réponse de type entité nommée *personne*) ; ou un rôle sémantique (l'amorce «Why do patients in Africa have an almost total lack of access to ARV drugs ?» attend une cause).

Nous nous sommes appuyé sur ce critère pour valider l'homogénéité sémantique des distracteurs. Pour cela, nous sommes parti d'un corpus annoté manuellement selon la conformité des types sémantiques des options au type attendu de l'amorce. Nous l'avons ensuite annoté automatiquement pour vérifier qu'il est possible de reconnaître automatiquement l'homogénéité sémantique en fonction du type attendu par l'amorce.

3.4.3.1 *Annotation manuelle*

Pour rédiger des QCM de qualité, il est nécessaire que les options soient sémantiquement conformes au type attendu par l'amorce, afin d'éviter d'avoir des distracteurs évidents pour des apprenants n'ayant pas assimilé les notions évaluées, ce qui nous a amené à proposer la notion de spécificité sémantique. L'annotation manuelle des distracteurs selon cette conformité détermine si les options correspondent au type attendu par l'amorce (déduit par l'annotateur, et non le résultat d'un module d'analyse de la question). Les catégories d'annotation sont les suivantes :

— **type conforme** : le type de l'option est conforme au type attendu de l'amorce le plus précis. Celui-ci est un type spécifique si l'amorce contient un type explicite (par exemple, le type spécifique attendu par l'amorce «Which president had the most children ?» est «président»). Sinon, il s'agit d'un type d'entité nommée. Si le type de la réponse n'est pas donné, mais seulement sa relation sémantique (causes, définitions...), l'option est considérée comme étant de type conforme si elle constitue un argument possible pour ce rôle ;

— **type non conforme :** inclut les options dont le type est différent du type attendu par l'amorce ;

— **conformité inconnue :** représente les options dont il est impossible d'évaluer la conformité par rapport au type attendu par l'amorce, c'est-à-dire les options pour lesquelles l'annotateur ne peut pas identifier le type ou pour lesquelles il est impossible d'identifier le type attendu de l'amorce (par exemple, «When using a file...»).

3.4.3.2 *Annotation automatique*

Pour annoter automatiquement les options, nous avons pris en considération tous les types d'entité nommée définis par Stanford Named Entity Recognizer : *Time*, *Location*, *Organization*, *Person*, *Money*, *Percent*, *Date*, *Duration*, *Ordinal*, *Set* et *Miscellaneous*. Nous avons comparé le type d'entité nommée des options au type attendu par l'amorce, reconnu par un analyseur automatique de questions (Ligozat, 2013). Cet analyseur donne des informations sur le type d'entité nommée, le type spécifique et le rôle sémantique attendu par l'amorce. Nous ne vérifions que le type d'entité nommée attendu par l'amorce.

3.4.3.3 *Évaluation des annotations*

Afin d'évaluer l'homogénéité sémantique des options du point de vue de leur conformité avec le type attendu par l'amorce, 609 options ont été annotées, sur les 838 options du corpus qcmValidation (73 %).

ANNOTATION MANUELLE Le tableau 13 montre la répartition des options selon les catégories sémantiques définies pour l'annotation.

	Nombre	Pourcentage
Type conforme	460	75,5 %
Type non conforme	26	4,3 %
Conformité inconnue	123	20,2 %
Total	609	100,0 %

TABLE 13 – Répartition des options selon leur conformité avec le type attendu de l'amorce

Environ trois quarts des options correspondent au type attendu par l'amorce. Cette observation montre que la conformité au type attendu par l'amorce est un critère d'homogénéité sémantique. Cependant, la conformité de 20 % des options n'a pas pu être identifiée : le type d'entité nommée ne peut pas être pris en considération pour caractériser ces options.

ANNOTATION AUTOMATIQUE Le tableau 14 montre les résultats de l'évaluation de l'annotation automatique des options selon leur conformité avec le type attendu par l'amorce. Afin de comparer l'annotation automatique à l'annotation manuelle, nous ne comparons pas les options qui ont été classifiées manuellement comme étant de «conformité inconnue» car elles ne fournissent pas une base pour évaluer l'annotation automatique.

	Précision	Rappel	F-mesure
Type conforme	0,97	0,64	0,77
Type non conforme	0,10	0,69	0,17

TABLE 14 – Résultats de l'évaluation de l'annotation automatique relative au type attendu par l'amorce

Nous observons que les options classifiées comme étant de «type conforme» sont relativement bien classifiées. Cependant, ce n'est pas le cas pour l'autre catégorie. Nous avons identifié trois raisons principales causant ces mauvais résultats : l'analyseur de questions échoue à reconnaître le type des amorces non interrogatives (par exemple, l'amorce «Yoshiko is in New York City because...» est reconnue comme une amorce attendant un *lieu*, alors qu'elle attend une raison). Les autres raisons sont dues à des erreurs d'étiquetage de l'identifeur d'entités nommées ou à des options qui ne sont pas des entités nommées. Les exemples 12 et 13 montrent ces deux phénomènes.

Amorce : For how long has Rebecca Lolosoli been working with MADRE ? (*type : durée*)
Option : since the late 1990s (*type correct : durée, type étiqueté par l'identifieur d'entités nommées : date*)

Exemple 12 – Item constitué d'une option mal étiquetée par l'identifieur d'entités nommées

Amorce : Where does Yoshiko's adventure begin ? (*type : lieu*)
Option : at the TeenSay offices (*type correct : lieu, mais non annoté par l'identifieur d'entités nommées car il ne s'agit pas d'une entité nommée*)

Exemple 13 – Item constitué d'une option non étiquetée par l'identifieur d'entités nommées

Dans cette analyse, nous avons montré que la conformité des options au type attendu par l'amorce est un bon moyen de reconnaître l'homogénéité sémantique des distracteurs. Cependant, une partie non négligeable des amorces sont de type sémantique inconnu. Cela montre la difficulté de reconnaître cette homogénéité. Il est donc préférable de s'appuyer sur d'autres critères pour estimer l'homogénéité sémantique des distracteurs selon leurs types d'entité nommée.

3.4.4 *Similarité des types d'entité nommée*

Le premier critère d'homogénéité sémantique que nous avons présenté est fondé sur la conformité des options au type attendu par l'amorce. Cependant, nous avons vu que l'homogénéité d'un distracteur peut être estimée à partir de la réponse. Nous présentons ici un critère d'homogénéité fondé sur la similarité des types d'entité nommée des distracteurs et de la réponse. Nous considérons que l'identité de type d'entité nommée de deux termes est théoriquement un critère nécessaire pour qu'ils soient sémantiquement homogènes.

Nous nous sommes appuyé sur ce critère pour valider la définition de l'homogénéité sémantique des distracteurs. Pour cela, nous sommes parti d'un corpus annoté manuellement selon la similarité des types d'entité nommée des distracteurs et de la réponse, puis nous avons effectué une annotation automatique de ce corpus selon la similarité des types d'entité nommées.

3.4.4.1 *Annotation manuelle*

Au niveau sémantique, les distracteurs ont un sens différent de celui de la réponse. Cependant, une certaine homogénéité sémantique peut être trouvée entre les distracteurs et la réponse selon leurs types d'entité nommée. L'annotation manuelle des distracteurs selon cette conformité détermine si les distracteurs ont le même type d'entité nommée que

la réponse. L'annotation s'est fondée sur la taxonomie des entités nommées du système de questions-réponses QALC (Ferret *et al.*, 2000) (figure 20).

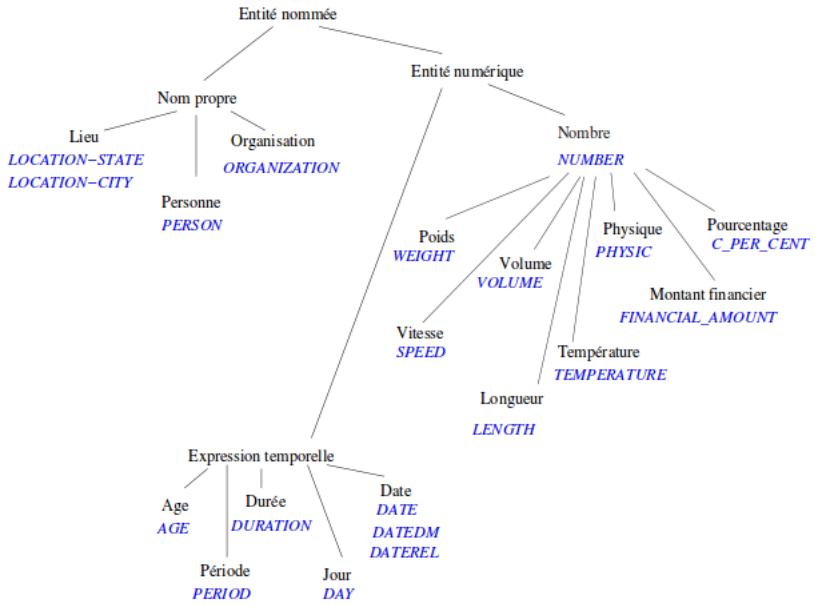


FIGURE 20 – Taxonomie de QALC

Cette taxonomie divise les types d'entités nommées en deux grandes catégories : les noms propres et les entités numériques. Les catégories d'annotation sont les suivantes :

— **type d'entité nommée identique :** représente les distracteurs partageant le même type d'entité nommée que la réponse ;

— **différent type d'entité nommée :** représente les distracteurs ne partageant pas le même type d'entité nommée que la réponse ;

— **pas une entité nommée :** représente les distracteurs qui ne sont pas des entités nommées.

3.4.4.2 *Annotation automatique*

Pour annoter automatiquement les distracteurs, nous avons pris en considération les mêmes types d'entité nommée que pour la validation de la conformité des options au type attendu de l'amorce (cf. section 3.4.3). Bien que les types soient différents de ceux sur lesquels s'est appuyé l'annotation manuelle, le but est similaire : comparer les types d'entité nommée des distracteurs et de la réponse. De plus, les deux ensembles de catégories d'entités nommées sont proches et peuvent être mis en correspondance.

3.4.4.3 *Évaluation*

Afin d'évaluer l'homogénéité sémantique des distracteurs du point de vue de la similarité du type d'entité nommée avec celui de la réponse, 484 distracteurs ont été annotés, sur les 650 distracteurs du corpus qcmValidation (74 %).

ANNOTATION MANUELLE Le tableau 15 montre la répartition des distracteurs selon les catégories sémantiques que nous avons définies.

	Nombre	**Pourcentage**
Type d'entité nommée identique	102	21,1 %
Type d'entité nommée différent	17	3,5 %
Pas une entité nommée	365	75,4 %
Total	484	100,0 %

TABLE 15 – Répartition des distracteurs selon leur homogénéité sémantique avec la réponse

Environ trois quarts des distracteurs ne sont pas de type d'entité nommée QALC. Quasiment tous les autres distracteurs ont le même type d'entité nommée que la réponse. Le tableau 16 montre la relation entre les types des options dans le cas des distracteurs ne partageant pas le type d'entité nommée de la réponse. Sur les 17 distracteurs concernés, 12 d'entre eux sont liés au type de la réponse (*pays* au lieu de *ville*, par exemple). Nous n'avons pas trouvé de distracteur dont le type d'entité nommée n'est pas lié à celui de la réponse.

	Nombre	**Pourcentage**
Hyperonymie	4	23,5 %
Hyponymie	8	47,1 %
Autre (différent type d'entité nommée	0	0,0 %
Autre (pas une entité nommée)	5	29,4 %
Total	17	100,0 %

TABLE 16 – Relations entre les distracteurs et les réponses de types d'entité nommée différents

Cette évaluation valide l'homogénéité sémantique des distracteurs de type entité nommée.

ANNOTATION AUTOMATIQUE Le tableau 17 montre les résultats de l'évaluation de l'annotation automatique des distracteurs selon le critère sémantique des entités nommées.

	Précision	**Rappel**	**F-mesure**
Type d'entité nommée identique	0,94	0,61	0,74
Type d'entité nommée différent	0,32	0,44	0,34
Pas une entité nommée	0,92	1,00	0,96

TABLE 17 – Résultats de l'évaluation de l'annotation sémantique automatique

Nous observons que la quasi-totalité des distracteurs qui ne sont pas du type entité nommée sont bien reconnus. De plus, une grande partie des distracteurs de même type d'entité nommée que la réponse sont identifiés et plus de la moitié des distracteurs annotés manuellement comme étant de «type d'entité nommée identique» sont bien reconnus. Les autres sont souvent reconnus comme n'étant «pas une entité nommée» car leur type d'entité nommée n'est pas une catégorie de Stanford Named Entity Recognizer ou sont mal étiquetés, comme le distracteur «The Methodist Church» qui n'est pas reconnu comme étant une organisation. Les distracteurs partageant un type d'entité nommée différent de la réponse ne sont pas très bien catégorisés : un tiers d'entre eux est reconnu comme étant de «type d'entité nommée identique» et un autre tiers est reconnu comme n'étant «pas une entité nommée».

Cette annotation automatique reproduit l'annotation manuelle, donc nous pouvons considérer qu'il est possible d'estimer la qualité des distracteurs suivant le type de l'entité nommée.

CATÉGORISATION DES COUPLES DE RÉPONSES-DISTRACTEURS SELON LEURS TYPES D'ENTITÉ NOMMÉE L'analyse précédente a montré qu'il existe une homogénéité sémantique entre les distracteurs et la réponse. Cependant, une partie des distracteurs n'ont pas le même type d'entité nommée que la réponse, bien qu'ils soient pertinents. L'analyse suivante a pour objectif de vérifier les cas où il existe une homogénéité sémantique même si le distracteur et la réponse ne sont pas de même type d'entité nommée.

Nous avons effectué une analyse poussée des distracteurs sur un nombre réduit de classes d'entités nommées (Location, Organization et Person). Pour cela, nous avons analysé le corpus qcmEN. Dans l'ensemble de ce corpus, 78 % des distracteurs sont de même type d'entité nommée que la réponse, ce qui confirme l'analyse précédente. Le tableau 18 montre les détails des combinaisons de types d'entité nommée comparés du corpus qcmEN.

Nous observons qu'il existe une proportion non négligeable des couples de réponses-distracteurs dont les types d'entité nommée sont différents, mais font partie de deux

Type EN 1	**Type EN 2**	**Nombre**	**Proportion**
lieu	lieu	115	38,5 %
personne	personne	76	25,4 %
organisation	organisation	27	9,0 %
lieu	organisation	16	5,4 %
personne	organisation	11	3,7 %
pas de type EN	*pas de type EN*	11	3,7 %
lieu	*pas de type EN*	10	3,3 %
personne	*pas de type EN*	8	2,7 %
lieu	personne	7	2,3 %
organisation	*pas de type EN*	7	2,3 %
lieu	*divers*	4	1,3 %
organisation	*divers*	2	0,7 %
divers	*divers*	2	0,7 %
durée	*pas de type EN*	2	0,7 %
personne	*divers*	1	0,3 %
		299	100,0 %

TABLE 18 – Quantités et proportions des combinaisons de types d'entités nommées entre réponses et distracteurs

des trois types d'entité nommée que nous prenons en compte : 11 % des couples de réponses-distracteurs font partie de cette catégorie. Bien que ces types d'entité nommée soient différents, il s'avère que les distracteurs concernés sont pertinents car ils ont les caractéristiques sémantiques requises pour répondre potentiellement à l'amorce associée, comme le montre l'exemple 14.

Amorce : Who withdrew France from NATO ?
Réponse : Charles de Gaulle (*type : Personne*)
Distracteur : Nicolas Sarkozy (*type : Personne*)
Distracteur : The European Union (*type : Organisation*)

Exemple 14 – Exemple d'item

Dans cet item, le distracteur «The European Union» n'a pas le même type d'entité nommée que la réponse. Cependant, ce distracteur est pertinent car il est possible de désigner une organisation ou un lieu comme une entité capable de décider («Paris a décidé d'envoyer des troupes au Mali»).

Les autres cas où les types d'entité nommée des distracteurs sont différents de celui de la réponse correspondent principalement à des entités mal annotées par Stanford Named Entity Recognizer (par exemple, «Washington, DC» a été étiqueté comme étant une organisation plutôt qu'un lieu).

Cela confirme l'homogénéité sémantique du point de vue des types d'entité nommée qui a été établie par l'annotation manuelle et montre qu'elle peut être vérifiée par des méthodes automatiques.

3.5 CONCLUSION

Dans l'objectif de répondre à notre problématique, c'est-à-dire l'évaluation automatique de la qualité d'un distracteur, nous nous fondons sur les propriétés syntaxiques et sémantiques communes entre le distracteur et la réponse. Pour être pertinents, le distracteur et la réponse doivent être syntaxiquement et sémantiquement homogènes, ce qui signifie que d'un côté, les distracteurs doivent partager une structure syntaxique commune à la réponse, et de l'autre côté, ils doivent avoir un sens proche de la réponse, mais ne doivent pas être similaires à celle-ci. Pour estimer l'homogénéité syntaxique entre un distracteur et la réponse, nous nous appuyons sur une comparaison de leurs arbres syntaxiques selon différentes mesures de similarité syntaxique. Pour estimer l'homogénéité sémantique entre un distracteur et la réponse, nous pouvons calculer plusieurs mesures de voisinage et de similarité sémantique fondées sur des représentations structurées. Cependant, la couverture de telles ressources est limitée. Pour compenser cette limitation, des mesures fondées sur des corpus peuvent également être calculées : fondées sur le fait que des termes sont similaires s'ils apparaissent dans des contextes similaires, elles s'appliquent

sur des corpus de textes non annotés, mais ne donnent pas la nature des relations entre distracteur et réponse.

Le degré de pertinence entre un distracteur et une réponse n'est pas absolu : un distracteur est plus pertinent que d'autres termes partageant des propriétés syntaxiques communes à la réponse. Pour estimer le degré de pertinence, nous proposons un modèle d'ordonnancement des candidats, filtrés par des critères d'homogénéité syntaxique, s'appuyant sur une combinaison de mesures de voisinage sémantique.

Nous avons également cherché à valider la possibilité de reconnaître automatiquement sur un corpus annoté manuellement. Les annotations manuelles ont montré qu'une grande partie des distracteurs sont homogènes d'un point de vue syntaxique et sémantique. Les annotations automatiques donnent des résultats corrects, et particulièrement l'annotation sémantique qui reconnaît 70 % des relations entre les entités nommées des distracteurs et de la réponse. L'annotation syntaxique automatique reconnaît la moitié des annotations manuelles représentant les relations entre les arbres syntaxiques des distracteurs et de la réponse, et particulièrement les distracteurs partageant la même syntaxe que la réponse. L'annotation liée au type attendu par les amorces reconnaît correctement les options dont les types d'entité nommées sont conformes au type attendu par l'amorce. Cependant, cette annotation est moins performante que l'annotation liée au type d'entité nommée des réponses. Pour évaluer l'homogénéité sémantique du point de vue des entités nommées, nous ne nous intéresserons qu'à la similarité entre les entités nommées des distracteurs et de la réponse, et pas de l'amorce.

Les chapitres suivants exposent la méthode d'évaluation automatique de la qualité des distracteurs que nous proposons, ainsi que les résultats de l'évaluation de cette méthode.

4

ÉVALUATION DE LA QUALITÉ DES DISTRACTEURS

Pour évaluer automatiquement la qualité des distracteurs, nous avons choisi un modèle d'ordonnancement de candidats fondé sur des critères d'homogénéité sémantique entre les candidats et la réponse. Dans ce chapitre, nous présentons chacun des traits utilisés par le modèle décrivant les candidats, la mise en œuvre du modèle, et ses résultats. Ce travail a donné lieu à deux publications (Pho *et al.*, 2015a,b).

L'analyse de corpus nous a conduit à approfondir la reconnaissance de l'homogénéité sémantique. Nous avons sélectionné des mesures fondées sur deux types de ressources : des connaissances sémantiques structurées, et des corpus de documents. Nous montrerons ici la manière dont nous les avons appliquées.

Ces mesures traduisent différents critères d'homogénéité sémantique : les mesures fondées sur les corpus calculent le degré de *voisinage sémantique* et les mesures fondées sur les connaissances structurées calculent le degré de *spécificité sémantique*.

Dans les sections suivantes, nous présentons les mesures fondées sur les connaissances structurées (sections 4.2 et 4.3) et les mesures fondées sur les corpus (section 4.4). Nous présentons ensuite le modèle d'ordonnancement à la section 4.5. Pour chacune de ces mesures, nous présentons et analysons ses résultats.

Les différentes mesures d'homogénéité sémantique et le modèle d'ordonnancement sont évalués en suivant une méthodologie similaire que nous présentons maintenant.

4.1 CONSTITUTION DU CORPUS D'APPRENTISSAGE ET MÉTHODOLOGIE D'ÉVALUATION

Pour apprendre et évaluer le modèle d'ordonnancement sémantique, ainsi que chacune des mesures d'homogénéité sémantique, nous avons constitué un jeu d'apprentissage composé d'exemples positifs et négatifs. Pour chaque item, les exemples positifs sont les distracteurs associés à l'item et les exemples négatifs sont des termes différents des distracteurs des items évalués et sélectionnés selon un critère d'homogénéité syntaxique avec la réponse. Nous appelons ces derniers des *non-distracteurs*. L'exemple 15 montre différents non-distracteurs d'un item.

Amorce : In what country is the Snow and Ice Data Center located ?
Réponse : United States
Distracteur 1 : Germany
Distracteur 2 : Denmark
Distracteur 3 : Mexico
Distracteur 4 : Bolivia
Non-distracteur 1 : North Polar
Non-distracteur 2 : Al Gore
Non-distracteur 3 : Australia

Exemple 15 – Exemple d'item composé de distracteurs et de non-distracteurs

Cependant, parmi les non-distracteurs sélectionnés, certains d'entre eux peuvent constituer des distracteurs pertinents, à l'instar du non-distracteur «Australia» dans l'exemple 15.

Notre objectif est d'apprendre un modèle d'ordonnancement capable de classer les distracteurs dans les premiers rangs, étant donné qu'ils devraient être plus homogènes à la réponse que les non-distracteurs.

Comme il a été indiqué au chapitre 3, les items que nous traitons sont associés à un document de référence à partir duquel les amorces sont conçues. La figure 21 montre les différentes étapes de l'évaluation automatique de la qualité des distracteurs. Une première étape consiste à annoter les corpus. Ensuite, les non-distracteurs sont extraits. La dernière étape consiste à apprendre le modèle à partir des candidats sélectionnés.

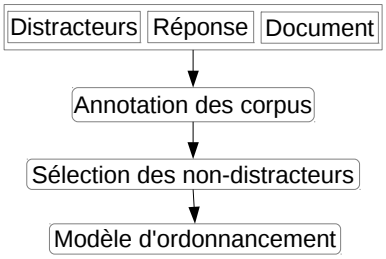


FIGURE 21 – Architecture montrant les différentes étapes de l'apprentissage du modèle

4.1.1 *Annotation des corpus*

Pour extraire les non-distracteurs et calculer les différentes mesures d'homogénéité, les candidats et les réponses doivent être annotés par des informations syntaxiques et sémantiques. L'annotation est de meilleure qualité si ces extraits sont analysés dans leur

contexte, c'est-à-dire le document de référence. Ainsi, nous effectuons quatre annotations du document dans l'ordre suivant :

1. une analyse syntaxique avec l'outil Stanford Parser ;
2. une annotation en entités nommées avec l'outil Stanford Named Entity Recognition. Cependant, certains termes ne sont pas annotés (à tort) par Stanford Named Entity Recognition. Nous nous appuyons sur l'analyse syntaxique de ces termes pour vérifier s'il s'agit de syntagmes nominaux. Si c'est le cas, nous identifions sa tête syntaxique et, si elle est reconnue comme étant une entité nommée par Stanford Named Entity Recognizer, ou elle appartient à une liste de déclencheurs d'entités nommées, nous attribuons le type d'entité nommée de la tête au terme ;
3. une annotation en types spécifiques pour trouver les entités associées à des entités de DBpédia (et, par extension, des pages de Wikipédia). Cette annotation est effectuée avec l'annotateur DBpedia Spotlight (Daiber *et al.*, 2013), qui associe des entités de DBpédia aux termes correspondants du document, et désambiguïse ces termes si nécessaire. Cependant, certains termes ne sont pas annotés (à tort) par DBpedia Spotlight. Nous associons ces termes à toutes les entités de DBpédia dont les titres correspondent à ces termes, donc sans désambiguïsation ;
4. une annotation en synsets de WordNet, visant à associer les termes (candidats et réponse) avec un ou plusieurs synsets de WordNet, comme suit :
   — si le terme apparaît dans WordNet, le terme est associé à ses synsets correspondants ;
   — si le terme n'apparaît pas dans WordNet et n'est pas une entité nommée, elle est associée aux synsets correspondants à sa tête syntaxique (par exemple, le chunk «the little cat» est associé aux synsets de WordNet de nom «cat»).

Les annotations des options sont extraites de leurs correspondances dans le document. Si une option n'apparaît pas dans le document, elle est annotée hors contexte de la même manière que l'est le document.

4.1.2 *Sélection des non-distracteurs*

Afin d'estimer le degré d'homogénéité des distracteurs d'un item, ceux-ci sont comparés à des non-distracteurs, sélectionnés selon deux méthodes différentes (chacune des méthodes est évaluée séparément) : la première méthode consiste à extraire les non-distracteurs à partir du document de référence de l'item (évaluation evalNDdocument), et la seconde consiste à sélectionner les options des autres items du corpus évalué (évaluation evalNDoption).

Les non-distracteurs sont sélectionnés s'ils sont syntaxiquement homogènes à la réponse. Si celle-ci est une entité nommée, les non-distracteurs sont les entités nommées

et/ou les syntagmes nominaux du document. Les entités nommées sont celles préalablement annotées et les syntagmes nominaux sont sélectionnés à partir des arbres de constituants des phrases du document ou des options des autres items, avec Tregex (Levy et Andrew, 2006), un outil permettant de sélectionner des nœuds d'arbres syntaxiques à partir de patrons. Si la réponse est un chunk et n'est pas une entité nommée, les non-distracteurs sont les chunks de même type que la réponse, ou de type syntaxique similaire [1] à celui de la réponse. Les chunks sont sélectionnés de la même manière que les syntagmes nominaux. Ces non-distracteurs sont annotés comme les options pour leur associer un type d'entité nommée, les entités de DBpédia et les synsets de WordNet.

Pour les deux évaluations, un dernier filtrage consiste à retirer les non-distracteurs similaires à une option, afin d'éviter les chevauchements sémantiques : deux termes sont considérés comme similaires s'ils sont associés aux mêmes entités de DBpédia ou s'ils réfèrent aux mêmes synsets dans WordNet (sauf si les synsets ne réfèrent qu'aux têtes des termes). Parmi les non-distracteurs sélectionnés, certains d'entre eux pourraient être assez pertinents pour être des distracteurs mais ne le sont pas car l'item contient assez de distracteurs, ou sont des distracteurs d'autres items. Dans nos travaux, nous traitons ces non-distracteurs comme des non-distracteurs normaux mais nous envisageons de faire une annotation manuelle des non-distracteurs afin d'écarter ces cas.

Dans chacun des corpus évalués, et chacune des évaluations proposées (evalNDdocument et evalNDoption), les non-distracteurs extraits sont différents. Le tableau 19 montre le nombre de distracteurs et de non-distracteurs extraits dans les corpus qcmEN et qcmNonEN, pour les évaluations evalNDdocument et evalNDoption.

	qcmEN	**qcmNonEN**
Distracteurs	299	545
Non-distracteurs (evalNDdocument)	7 583	42 396
Non-distracteurs (evalNDoption)	27 390	69 558

TABLE 19 – Nombre de distracteurs et de non-distracteurs totaux des corpus qcmEN et qcmNonEN et pour les évaluations evalNDdocument et evalNDoption

Pour chacune des mesures évaluées (excepté le modèle), nous n'indiquerons que les résultats de l'évaluation evalNDdocument. En effet, les observations effectuées sur les deux évaluations sont similaires, même si les résultats peuvent être différents. Les résultats de l'évaluation evalNDoption pourront être connectés aux résultats de l'évaluation evalNDdocument en annexe.

1. Nous considérons similaires les syntagmes nominaux, les chunks adjectivaux et les chunks verbaux dont la tête syntaxique est un verbe conjugué au participe présent ou passé

Pour apprendre le modèle d'ordonnancement, nous gardons tous les candidats des corpus. Contrairement à un modèle de classification n-aire, nous n'avons pas besoin d'équilibrer le nombre d'exemple de chacune des classes pour éviter un biais lors de l'apprentissage.

4.1.3 *Classement des candidats et évaluation*

Le modèle d'ordonnancement, et chacun de ses critères d'homogénéité sémantique, sont évalués selon leur capacité à ordonner correctement les candidats, c'est-à-dire à classer les distracteurs dans les premiers rangs. Pour vérifier cela, nous avons effectué quatre analyses.

1. Afin de vérifier la portée de ces mesures, nous calculons la proportion de distracteurs et de non-distracteurs disponibles dans la ressource associée à la mesure calculée. Nous indiquons notamment la proportion de *couples de réponses-distracteurs* et de *couples de réponses-non-distracteurs*, c'est-à-dire des couples dont la réponse et le candidat sont disponibles dans la ressource. Les analyses suivantes ne s'effectuent que sur les candidats de ces couples.
2. Nous évaluons l'ordonnancement des candidats à partir de plusieurs mesures d'évaluation (plus de détails dans la section 4.1.3.1).
3. Nous analysons la répartition des distracteurs et des non-distracteurs selon les scores d'homogénéité. Cette analyse permet de vérifier si une séparation entre les distracteurs et les non-distracteurs est possible, et que les tendances entre les distracteurs et les non-distracteurs sont différentes.

4.1.3.1 *Mesures d'évaluation*

Pour évaluer l'ordonnancement des candidats effectué par chacune des mesures d'homogénéité sémantique, nous considérons que les distracteurs sont sémantiquement plus proches de la réponse que les non-distracteurs et, par conséquent, devraient avoir un meilleur rang dans le classement. *Nous calculons ces mesures d'évaluation sur la totalité des corpus, et sur la partie des corpus couverte par la ressource associée à la mesure.* Afin d'évaluer cela, nous calculons la précision (équation 20) et le rappel (équation 21) moyens au rang n en fonction du nombre de distracteurs de l'item, ainsi que la f-mesure (équation 22).

$$PM = \frac{\sum_{i}^{nbI} P_{i,nbD}}{nbI} \tag{20}$$

$$RM = \frac{\sum_{i}^{nbI} R_{i,nbD}}{nbI} \tag{21}$$

$$FM = 2\frac{PM \times PR}{PM + PR} \tag{22}$$

où nbI est le nombre d'items du corpus, nbD le nombre de distracteurs de l'item évalué, et $P_{i,nbD}$ et $R_{i,nbD}$ sont la précision (équation 23) et le rappel (équation 24) de l'item i.

$$P_{i,nbD} = \frac{\#\text{ distracteurs de rang} \leqslant nbD}{\#\text{ candidats de rang} \leqslant nbD} \tag{23}$$

$$R_{i,nbD} = \frac{\#\text{ distracteurs de rang} \leqslant nbD}{nbD} \tag{24}$$

Ces mesures sont intéressantes pour vérifier que les distracteurs sont bien classés dans les premiers rangs. Cependant, elles ne prennent pas en considération les distracteurs obtenant un rang correct mais moins homogènes que certains non-distracteurs (notamment les non-distracteurs assez pertinents pour constituer des distracteurs). Une mesure prenant en compte ce critère est la *mean average precision* (équation 25).

$$MAP = \frac{\sum_{i}^{nbI} Pmoy_{i,nbD}}{nbI} \tag{25}$$

où $Pmoy_{i,nbD}$ est la précision moyenne de l'item i.

$$Pmoy_{i,nbD} = \frac{\sum_{d}^{nbD} \frac{\text{rang relatif de } d}{\text{rang de } d}}{nbD} \tag{26}$$

où d est un distracteur, son rang relatif est le rang de d selon l'ordonnancement des distracteurs (et pas des non-distracteurs) de i, et son rang absolu est le rang de d selon l'ordonnancement de tous les candidats de i.

Le rappel moyen permet de vérifier la proportion de distracteurs faisant partie des candidats de rang $\leqslant nbD$, tandis que la précision moyenne permet de vérifier la proportion de distracteurs de rang $\leqslant nbD$ et ne partageant pas leurs rangs avec des non-distracteurs. La *mean average precision* permet de vérifier le classement moyen des distracteurs. Une mesure obtenant un *rappel moyen* et une *MAP* élevés, et une *précision moyenne* faible indique donc qu'une grande partie des distracteurs sont classés dans les premiers rangs, mais qu'un grand nombre de non-distracteurs sont également classés dans les premiers rangs. Cela peut montrer que la mesure peut être efficace pour indiquer des candidats non sémantiquement homogènes, mais peut aussi traduire une faible couverture de la ressource associée à la mesure.

Les exemples 16 et 17 montrent une application des mesures présentées dans cette section.

Amorce : Who is the founder of the SING campaign ?
Réponse : Annie Lennox
Distracteur 1 (d_1) : Nelson Mandela
(*score ESA : 0,018, rang : 2, rang relatif : 2*)
Distracteur 2 (d_2) : Youssou N'Dour (*score ESA : 0,014, rang : 3, rang relatif : 3*)
Distracteur 3 (d_3) : Michel Sidibe (*score ESA : 0,020, rang : 1, rang relatif : 1*)
Distracteur 4 (d_4) : Zackie Achmat (*score ESA : 0, rang : 5, rang relatif : 4*)
Non-distracteur 1 (nd_1) : Aaron Motsoaledi
(*score ESA : 0,012, rang : 4*)
Non-distracteur 2 (nd_2) : Avelile (*score ESA : 0, rang : 5*)

Exemple 16 – Exemple d'item composé de distracteurs et de non-distracteurs

Cet item contient 4 distracteurs, donc $nbD = 4$. Les candidats de rang inférieur à 4 sont d_3, d_1, d_2 et nd_1 (3 distracteurs font partie de cet ensemble). Le rappel et la précision sont donc de $\frac{3}{4} = 0,75$. La MAP de cet item est de $\frac{\frac{2}{2}+\frac{3}{3}+\frac{1}{1}+\frac{4}{5}}{4} = \frac{3,8}{4} = 0,95$.

Amorce : Who was the first president to have a dog in the White House ?
Réponse : George Washington
Distracteur 1 (d_1) : Bill Clinton (*score EN : 1, rang : 1, rang relatif : 1*)
Distracteur 2 (d_2) : John F. Kennedy (*score EN : 1, rang : 1, rang relatif : 1*)
Non-distracteur 1 (nd_1) : Barry H. Landau (*score EN : 1, rang : 1*)
Non-distracteur 2 (nd_2) : Betty Curie (*score EN : 1, rang : 1*)
Non-distracteur 3 (nd_3) : Obama (*score EN : 1, rang : 1*)

Exemple 17 – Exemple d'item composé de distracteurs et de non-distracteurs

Cet item contient 2 distracteurs, donc $nbD = 2$. Il ne contient pas de candidat de rang inférieur à 2. Le rappel est donc de $\frac{2}{2} = 1$, la précision est de $\frac{2}{5} = 0,4$ et la MAP est de $\frac{\frac{1}{1}+\frac{1}{1}}{2} = \frac{2}{2} = 1$. Les distracteurs sont reconnus, mais les non-distracteurs sont également reconnus comme des distracteurs.

Pour chacune des mesures d'homogénéité sémantique, nous indiquons également le pourcentage de **non-distracteurs mal reconnus** des parties des corpus couvertes par la ressource associée à la mesure (dans le cas de la mesure de similarité des types d'entité nommée et du modèle, l'évaluation s'effectue sur l'ensemble des corpus). Un non-distracteur nd_i appartenant à l'item i est mal reconnu si $rang(nd_i) \leqslant nbD_i$ ($rang(nd_i)$ est le rang de nd_i parmi les candidats de i et nbD_i est le nombre de distracteurs de i).

Les mesures que nous avons présentées ont pour but d'évaluer l'ordonnancement des candidats effectué par le modèle proposé, ainsi que de chacun de ses critères d'homogénéité sémantique. Elles permettent de vérifier la position des distracteurs parmi les candidats. Le modèle, et les critères d'homogénéité sémantique, sont performants s'ils classent les distracteurs dans les premiers rangs. C'est pourquoi nous avons étudié la ca-

pacité de chaque mesure à séparer distracteurs et non-distracteurs. Afin d'en donner une vision globale, nous avons regroupé les figures (cf. figures 25).

4.2 MESURES DE VOISINAGE SÉMANTIQUE FONDÉES SUR LES TYPES SÉMANTIQUES

Les mesures fondées sur les types sémantiques calculent le voisinage sémantique des termes selon le degré de similarité de leurs types sémantiques. Ces types peuvent être généraux, comme les types d'entité nommée, ou spécifiques et hiérarchisés, comme les types fournis par la taxonomie de DBpédia. Dans cette section, nous présentons une mesure fondée sur la similarité des types d'entité nommée (section 4.2.1), puis nous présentons une mesure fondée sur la similarité des types sémantiques provenant de DBpédia (section 4.2.2).

4.2.1 *Similarité des types d'entité nommée*

Le premier critère d'homogénéité sémantique que nous présentons est fondé sur une annotation des textes en entités nommées, c'est-à-dire classiquement les noms de lieux, personnes et organisations. Nous considérons que l'identité de type d'entité nommées de deux termes est théoriquement un critère nécessaire pour qu'ils soient sémantiquement homogènes. Ce critère n'est pas suffisant car les types d'entité nommée choisis sont des catégories très générales, et qu'il convient également d'exclure les termes similaires.

Afin de mesurer le critère d'homogénéité sémantique des candidats selon leurs types d'entité nommée, nous considérons trois grandes catégories : *lieu*, *organisation* et *personne*. Pour comparer le type d'entité nommée de deux termes, nous utilisons la mesure binaire $\mathrm{meme\_type\_EN}$ (équation 27) qui indique simplement si les termes sont de même type d'entité nommée ou non.

$$\mathrm{meme\_type\_EN}(t_1, t_2) = \begin{cases} 1 \text{ si} & \mathrm{EN}(t_1) = \mathrm{EN}(t_2) \\ & \wedge t_1 \text{ est une EN} \\ & \wedge t_2 \text{ est une EN} \\ 0 \text{ sinon} & \end{cases} \tag{27}$$

où t_1 et t_2 sont deux termes et $\mathrm{EN}(t)$ est le type d'entité nommée du terme t.

4.2.2 *Similarité des types sémantiques provenant de DBpédia*

En plus de mesurer la similarité de types d'entité nommée généraux, nous souhaitons comparer les types sémantiques des termes à un niveau de granularité plus fin : des types plus spécifiques permettent de vérifier plus précisément l'homogénéité sémantique. Ainsi,

dans l'exemple 18, les options sont toutes des pays asiatiques, ce qui donne une indication plus précise de leur homogénéité sémantique que de vérifier simplement qu'elles sont des lieux.

Amorce : De quel pays est originaire le kimchi ?
Réponse : Corée
Distracteur : Japon
Distracteur : Chine
Distracteur : Mongolie

Exemple 18 – Item à choix multiples

Cependant, tandis que les types d'entité nommée peuvent être reconnus indépendamment d'une ressource, les types spécifiques doivent être reconnus à partir d'une taxonomie hiérarchique. Pour cela, nous utilisons la ressource DBpédia (Auer *et al.*, 2007) que nous avons décrite au chapitre 2. Cette ressource a l'avantage d'avoir une large couverture en domaine ouvert.

Pour calculer l'homogénéité sémantique entre deux termes fondée sur leur type DBpédia et leur position dans leur taxonomie, nous utilisons la mesure de Wu et Palmer (cf. chapitre 2) sur les types DBpédia des termes. Cette mesure est fondée sur le chemin le plus court entre les deux types comparés, pondérés par leur profondeur dans DBpédia. Ainsi, deux types spécifiques de parent commun obtiennent un meilleur score que deux types moins spécifiques de parent commun.

4.2.3 *Évaluation*

4.2.3.1 *Couverture des candidats dans DBpédia*

Le tableau 20 donne les proportions de réponses, distracteurs, non-distracteurs, couples de réponses-distracteurs et couples de réponses-non-distracteurs des corpus qcmEN et qcmNonEN apparaissant dans DBpédia et ayant un type DBpédia, pour les évaluations evalNDdocument et evalNDoption.

Nous observons que 20 % des distracteurs et 60 % des non-distracteurs provenant des documents de référence du corpus qcmEN ne réfèrent pas à une entité de DBpédia ou ne sont pas typés. Ces candidats sont principalement des personnes et des organisations non célèbres, donc non répertoriées par DBpédia (par exemple, le village néolithique «Skerrabra», situé dans l'île Mainland de l'archipel des Orcades). Bien que la majorité des distracteurs soit couverte par DBpédia, il reste néanmoins une partie non négligeable de candidats dont nous ne pouvons définir de type spécifique.

Concernant les chunks non entité nommée, un grand nombre de réponses et de candidats du corpus qcmNonEN ne sont pas présents dans DBpédia. En effet, un grand nombre de ces chunks ne sont pas des entités comme par exemple, le non-distracteur «*the*

	qcmEN	qcmNonEN
réponses	86 (83,5 %)	39 (19,9 %)
distracteurs	240 (80,3 %)	87 (16,0 %)
non-distracteurs (evalNDdocument)	2952 (38,9 %)	5591 (13,2 %)
non-distracteurs (evalNDoption)	21717 (79,3 %)	12986 (18,7 %)
couples R-D	208 (69,6 %)	29 (5,3 %)
couples R-ND (evalNDdocument)	2487 (32,8 %)	1299 (3,1 %)
couples R-ND (evalNDoption)	18274 (66,7 %)	2779 (4,0 %)

TABLE 20 – Couverture de DBpédia (entités dont un type DBpédia est attribué)

national food». DBpédia ne contenant que des entités, cette ressource ne couvre que des syntagmes nominaux.

Cela montre que la prise en compte de la spécificité du type ne peut être qu'une information indicative et non absolue, en particulier pour les chunks non entité nommée qui sont peu présents dans DBpédia.

4.2.3.2 *Calcul des mesures d'évaluation*

Le tableau 21 et la figure 22 montrent les résultats de l'évaluation des mesures fondées sur la similarité des types.

	meme_type_EN	*wup* (types DBpédia)			
	qcmEN	**qcmEN**		**qcmNonEN**	
		Couverture	**Corpus**	**Couverture**	**Corpus**
RM	0,83	0,66	0,68	0,68	0,93
PM	0,13	0,38	0,30	0,37	0,07
FM	0,23	0,48	0,41	0,48	0,13
MAP	0,85	0,74	0,73	0,76	0,92

TABLE 21 – Résultats de l'évaluation des mesures fondées sur les types

Nous observons que le rappel moyen est considérablement plus élevé que la précision moyenne. Cela est principalement dû au fait qu'un grand nombre de distracteurs ont le même score que les non-distracteurs, comme le montre l'exemple 19.

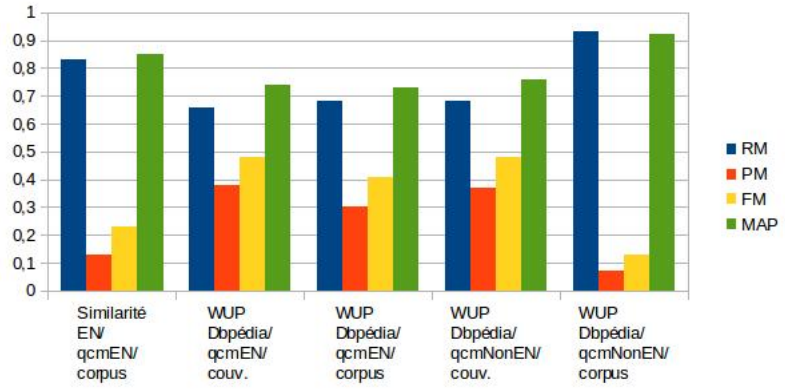


FIGURE 22 – Comparaison des résultats de l'évaluation des mesures fondées sur les types

Amorce : What is Nelson Mandela's country of origin ?
Réponse : South Africa
(*type EN : lieu, type spécifique : Country*)
Distracteur : Namibia
(*type EN : lieu, type spécifique : Country,* meme_type_EN $= 1$, *wup (types DBpédia)* $= 1$
Non-distracteur : Hong Kong
(*type EN : lieu, type spécifique : Country,* meme_type_EN $= 1$, *wup (types DBpédia)* $= 1$

Exemple 19 – Exemple d'item contenant une réponse et un distracteur, et un non-distracteur

Les non-distracteurs reconnus comme étant des distracteurs par la mesure meme_type_EN sont principalement des non-distracteurs de même type d'entité nommée que la réponse. Une partie de ces non-distracteurs n'a pas le même type d'entité nommée que la réponse, mais a été sélectionnée car ces non-distracteurs sont dans les premiers rangs des candidats. Une autre partie constitue les distracteurs de type spécifique proche de celui de la réponse (ville – pays). Les autres non-distracteurs sont assez pertinents pour être des distracteurs. La prise en compte du type d'EN est un critère pertinent, mais non suffisant : il permet de donner une indication sur l'homogénéité sémantique mais ne suffit pas car elle s'appuie sur des types généraux.

Les non-distracteurs des corpus qcmEN et qcmNonEN reconnus comme étant des distracteurs par la mesure fondée sur les types DBpédia sont répartis en deux catégories : les non-distracteurs assez pertinents pour remplacer des distracteurs et les non-distracteurs dont le type DBpédia est plus ou aussi similaire à celui de la réponse que les distracteurs.

La mesure fondée sur les types DBpédia donne une meilleure précision moyenne que la mesure meme_type_EN. Cela signifie que les types DBpédia sont plus intéressant que les entités nommées pour reconnaître les distracteurs. Cependant, la ressource DBpédia ne couvre pas tous les termes, et notamment les chunks non entité nommée dont la couverture est très faible.

4.2.3.3 *Répartition des candidats selon le score de voisinage sémantique fondé sur les types DBpédia*

Les figures 25a et 25b montrent la répartition des candidats selon leurs scores de voisinage sémantique fondés sur la similarité des types de DBpédia.

D'après ces figures, entre un tiers et la moitié des distracteurs sont de même type spécifique que la réponse ($wup = 1$), tandis que 10 % des non-distracteurs ont cette caractéristique. Nous observons également qu'environ 40 % des distracteurs et des non-distracteurs de type entité nommée ont un type spécifique sémantiquement proche de la réponse, mais non identique ($0 < wup < 1$). La seconde moitié des non-distracteurs de type entité nommée sont de type différent de celui de la réponse ($wup = 0$).

Dans le cas des candidats de type chunk non entité nommée, la plupart des non-distracteurs ont un type DBpédia différent de la réponse, tandis que les distracteurs sont équitablement répartis selon les scores de *wup*.

Nous pouvons conclure qu'un candidat de même type spécifique que la réponse a une forte probabilité d'être un distracteur et qu'un candidat de type différent de la réponse a une forte probabilité d'être un non-distracteur. Cependant, il est plus difficile de dissocier les distracteurs de type proche de celui de la réponse des autres candidats de même propriété.

Cette analyse nous montre que les types d'entité nommée généraux ou spécifiques, issus de bases de connaissances, peuvent donner une indication sur la pertinence des distracteurs, mais que cette caractéristique n'est pas absolue pour évaluer leur homogénéité.

4.3 MESURES DE VOISINAGE SÉMANTIQUE FONDÉES SUR WORDNET

Les mesures précédentes sont fondées sur la similarité des types sémantiques des termes. Dans cette section, nous présentons des mesures de voisinage sémantique fondées sur le sens des termes et les relations qui les lient.

Pour mesurer l'homogénéité sémantique sur tout type de termes et particulièrement les termes qui ne sont pas des entités nommées, nous utilisons des mesures définies pour WordNet. Nous utilisons les quatre mesures sélectionnées par Mitkov *et al.* (2009) dans leur travail de sélection automatique des distracteurs et qui ont été présentées à la section 2.4 : la mesure de recoupement étendu de gloses ; la mesure de Leacock et Chodorow fondée sur le plus court chemin entre les synsets ; les mesures de Jiang et Conrath, et de

Lin, fondées sur le contenu informationnel. Pour calculer le contenu informationnel des synsets, nous avons utilisé le corpus par défaut de WordNet, SemCor, un sous-ensemble du Brown Corpus en langue anglaise (échantillon provenant de sources comme des articles de journaux et de domaines variés).

Les termes pouvant avoir plusieurs sens, nous les associons à plusieurs synsets. Ainsi, pour calculer le voisinage sémantique entre deux termes, nous calculons les mesures sur toutes les paires de synsets associées aux termes et nous gardons le score maximal.

Ces mesures se complètent car elles calculent le score de voisinage sémantique selon différents critères : tandis que la mesure de Leacock et Chodorow se fonde uniquement sur les relations sémantiques explicites entre les synsets, la mesure de recoupement étendu de gloses se fonde sur la proximité textuelle des gloses des synsets et les mesures de Jiang et Conrath et Lin se fondent sur un corpus de textes pour comparer l'importance (représenté par la fréquence d'apparition) des synsets.

4.3.1 *Évaluation*

4.3.1.1 *Couverture des candidats dans WordNet et le corpus de fréquences*

Pour analyser les mesures fondées sur WordNet, nous avons calculé la proportion de réponses, distracteurs, non-distracteurs, couples de réponses-distracteurs et couples de réponses-non-distracteurs apparaissant dans WordNet et dans le corpus de fréquences associé à WordNet (tableau 22). Ces dernières mesures sont le fondement des mesures de Jiang et Conrath et Lin en plus de la taxonomie de WordNet.

	WordNet		**Corpus de fréquences**	
	qcmEN	**qcmNonEN**	**qcmEN**	**qcmNonEN**
réponse	54 (52,4 %)	159 (81,1 %)	34 (33,0 %)	112 (57,1 %)
distracteurs	160 (53,5 %)	423 (77,6 %)	96 (32,1 %)	264 (48,4 %)
n.-d. (evalNDd.)	2726 (35,9 %)	29942 (70,6 %)	1894 (25,0 %)	23477 (55,4 %)
n.-d. (evalNDo.)	13168 (48,1 %)	52128 (74,9 %)	7069 (25,8 %)	38246 (55,0 %)
c. R-D	129 (43,1 %)	378 (69,4 %)	70 (23,4 %)	199 (36,5 %)
c. R-ND (evalNDd.)	1220 (16,1 %)	25321 (59,7 %)	512 (6,8 %)	14442 (34,1 %)
c. R-ND (evalNDo.)	6686 (24,4 %)	41499 (59,6 %)	2161 (7,9 %)	23853 (34,2 %)

TABLE 22 – Couverture de WordNet

D'après ce tableau, nous observons que la moitié des options et des non-distracteurs extraits pour evalNDoption du corpus qcmEN apparaissent dans WordNet, tandis que 30 % des entités du corpus apparaissent dans le corpus de fréquences. En revanche, la

couverture de WordNet est considérablement plus grande sur les options de type chunk non entité nommée (plus des trois quarts des options couvertes par WordNet), car celles-ci sont généralement constitués de mots ou termes du dictionnaire. WordNet couvre tout type de chunk. Cependant, le corpus de fréquence associé à WordNet offre une couverture moins large que WordNet. Bien que WordNet couvre tout type de chunk, la plupart des mesures fondées sur WordNet se calculent sur les syntagmes nominaux et les chunks verbaux car les adjectifs et les adverbes ne sont pas hiérarchisés. La seule exception est la mesure de recoupement étendu de gloses, qui compare les gloses des synsets comparés. De plus, les taxonomies des noms et des verbes ne sont pas liées dans WordNet. La comparaison de noms et de verbes est donc impossible en prenant en considération les relations de WordNet.

Cette analyse montre que les mesures de voisinage sémantique fondées sur WordNet s'effectueront particulièrement sur des termes non entités nommés, contrairement à DBpédia qui offre une couverture plus large aux termes de type entité nommée.

4.3.1.2 *Calcul des mesures d'évaluation*

Les tableaux 23 et 24 montrent les résultats de l'évaluation des mesures fondées sur WordNet. Les figures 23 et 24 montrent une visualisation graphique des résultats de l'évaluation de ces mesures.

	reg				lch			
	qcmEN		**qcmNonEN**		**qcmEN**		**qcmNonEN**	
	Couv.	**Corpus**	**Couv.**	**Corpus**	**Couv.**	**Corpus**	**Couv.**	**Corpus**
RM	0,49	0,68	0,32	0,44	0,55	0,70	0,34	0,45
PM	0,49	0,22	0,14	0,14	0,45	0,21	0,16	0,13
FM	0,49	0,34	0,23	0,21	0,49	0,33	0,22	0,20
MAP	0,57	0,73	0,35	0,46	0,63	0,75	0,38	0,48

TABLE 23 – Résultats de l'évaluation des mesures de recoupement étendu de gloses (reg) et de Leacock et Chodorow (lch)

Les rappels et précisions moyens, ainsi que les f-mesures de ces mesures (excepté pour celle de Leacock et Chodorow) montrent que les mesures de voisinage sémantique fondées sur WordNet ne se limitent pas à un filtrage des candidats. Le cas de la mesure de Leacock et Chodorow est différent : étant donné qu'elle calcule le plus court chemin entre deux concepts, une partie des non-distracteurs obtiennent le même score que des distracteurs car ces candidats ont des mêmes ancêtres communs.

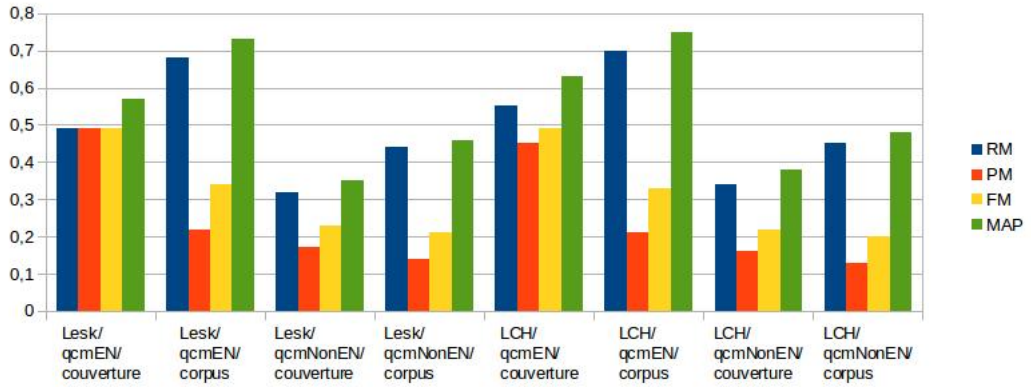


FIGURE 23 – Comparaison des résultats de l'évaluation des mesures de recoupement étendu de gloses (reg) et de Leacock et Chodorow (lch)

	jcn				lin			
	qcmEN		**qcmNonEN**		**qcmEN**		**qcmNonEN**	
	Couv.	**Corpus**	**Couv.**	**Corpus**	**Couv.**	**Corpus**	**Couv.**	**Corpus**
RM	0,48	0,80	0,23	0,54	0,58	0,81	0,24	0,55
PM	0,45	0,16	0,21	0,12	0,55	0,17	0,21	0,11
FM	0,47	0,26	0,22	0,19	0,57	0,28	0,23	0,18
MAP	0,59	0,82	0,31	0,57	0,64	0,83	0,32	0,57

TABLE 24 – Résultats de l'évaluation des mesures de Jiang et Conrath (jcn) et de Lin (lin)

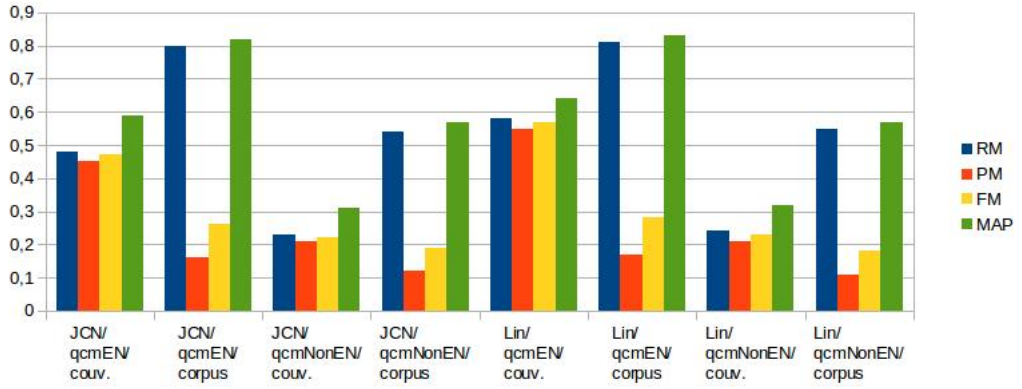


FIGURE 24 – Comparaison des résultats de l'évaluation des mesures de Jiang et Conrath (jcn) et de Lin (lin)

Nous observons que ces mesures sont plus performantes sur les entités nommées que sur les chunks non entité nommée. La raison principale est que, parmi les réponses de type chunk non entité nommée, une partie d'entre eux font référence à plusieurs concepts de WordNet, et sont sémantiquement proches de non-distracteurs qui ne sont pas forcément pertinents. Cependant, la couverture de WordNet est faible sur les entités nommées.

4.3.1.3 *Répartition des candidats selon les scores de voisinage sémantique fondés sur WordNet*

Les figures 25c, 25d, 25e, 25f, 25g, 25h, 25i et 25j montrent la répartition des candidats selon leurs scores de voisinage sémantique fondés sur WordNet.

Nous observons que les scores de Leacock et Chodorow, Jiang et Conrath et Lin distinguent nettement les distracteurs des non-distracteurs de type entité nommée. Cependant, cette distinction n'est pas établie pour les candidats de type chunk non entité nommée. Les répartitions des candidats sont similaires, qu'il s'agisse des distracteurs et des non-distracteurs.

Afin d'avoir également des mesures s'appliquant à tout type de termes sans la limite de leur présence ou non dans une ressource sémantique structurée, nous avons sélectionné des mesures pouvant être calculées sur de grands corpus. Cette famille de mesures ne prend pas en considération les relations sémantiques explicites, et sont dédiées à l'estimation du voisinage sémantique à l'instar des mesures précédentes concernant la mesure du voisinage sémantique.

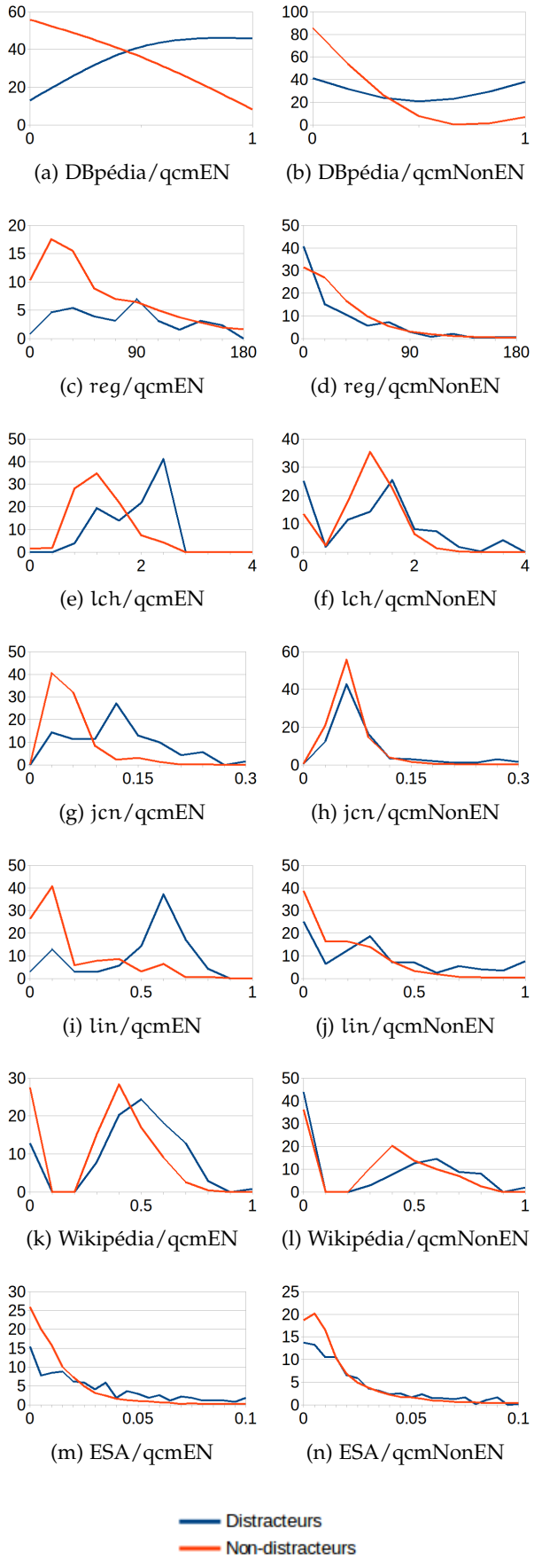


FIGURE 25 – Répartition des candidats en fonction des scores d'homogénéité sémantique. Les abscisses représentent le score d'homogénéité sémantique et les ordonnées représentent la proportion de distracteurs et de non-distracteurs du corpus

4.4 MESURES DE VOISINAGE SÉMANTIQUE FONDÉES SUR LES CORPUS

Les mesures fondées sur les corpus s'appuient sur de vastes corpus de documents pour estimer le degré de voisinage sémantique de deux termes. Les mesures que nous avons sélectionnées s'appuient sur différentes représentations des termes comparés : la mesure fondée sur la comparaison de liens de pages Wikipédia (section 4.4.1) représente un terme par l'ensemble de liens entrants et sortants de la page Wikipédia de référence, et la mesure fondée sur l'ESA (section 4.4.2) représente un terme par ses fréquences dans les documents du corpus.

4.4.1 *Comparaison des liens de pages Wikipédia*

Une représentation contextuelle possible d'un terme est l'ensemble des liens entrants et sortants associés à une page Wikipédia. Nous considérons les pages dont le titre correspond au terme de la réponse ou d'un candidat. Les liens entrants et sortants représentent les pages de Wikipédia associées au corps d'une page de Wikipédia. La mesure que nous calculons est fondée sur la similarité de ce liens entre les pages des termes comparés (cf. chapitre 2). L'outil Wikipedia Miner (Milne et Witten, 2013) effectue ce calcul à partir de dumps de Wikipédia.

4.4.2 *Analyse Sémantique Explicite*

Une autre représentation contextuelle des termes est leur distribution à travers les documents d'un corpus. Deux termes dont les distributions (c'est-à-dire les fréquences d'apparition) sont proches dans les mêmes documents ont une forte probabilité d'être sémantiquement voisins. Afin de comparer les distributions de deux termes, nous calculons une mesure fondée sur l'Analyse Sémantique Explicite (Gabrilovich et Markovitch, 2007) (cf. chapitre 2). Le corpus de documents sur lequel nous nous appuyons est Wikipédia. Pour calculer cette mesure, nous utilisons l'outil ESAlib [2].

Nous avons également évalué une autre mesure de voisinage sémantique fondée sur les corpus : la mesure fondée sur l'Analyse Sémantique Latente (Landauer et Dumais, 1997) (cf. chapitre 2), à partir de laquelle les contextes des candidats et de la réponse étaient calculés. Les contextes du candidat sont les passages (paragraphes ou phrases) du document de référence dans lesquels il apparaît, et le contexte de la réponse est un sac de mots contenant les mots de l'amorce et de la réponse. Nous n'avons pas retenu cette approche car l'évaluation a révélé une plus faible performance que d'autres mesures fondées sur les corpus comme la mesure fondée sur l'ESA. Néanmoins, dans de futurs travaux, nous comptons améliorer cette approche pour prendre en considération les passages du document de référence et l'amorce pour estimer l'homogénéité sémantique. L'évaluation

2. http://ticcky.github.io/esalib/

de cette mesure se trouve en annexe. Elle a été étudiée lors d'un séjour de recherche au CENTAL sous l'encadrement de Cédrick Fairon et Thomas François.

4.4.3 *Évaluation*

4.4.3.1 *Couverture des candidats*

Le tableau 25 donne les proportions de réponses, distracteurs, non-distracteurs, couples de réponses-distracteurs et couples de réponses-non-distracteurs des corpus qcmEN et qcmNonEN apparaissant dans Wikipédia et dans le corps de ses pages.

	Pages Wikipédia		**Corps pages Wikipédia**	
	qcmEN	**qcmNonEN**	**qcmEN**	**qcmNonEN**
réponses	94 (91,3 %)	129 (65,8 %)	96 (93,2 %)	175 (89,3 %)
distracteurs	261 (87,3 %)	344 (63,1 %)	292 (97,7 %)	505 (92,7 %)
n.-d. (evalNDd.)	4202 (55,4 %)	8945 (21,1 %)	7307 (96,4 %)	38207 (90,1 %)
n.-d. (evalNDo.)	23583 (86,1 %)	44893 (64,5 %)	26253 (95,8 %)	66484 (95,5 %)
c. R-D	242 (80,9 %)	249 (45,7 %)	272 (91,0 %)	474 (87,0 %)
c. R-ND (evalNDd.)	3818 (50,3 %)	5994 (14,1 %)	6261 (82,6 %)	34752 (82,0 %)
c. R-ND (evalNDo.)	21464 (78,3 %)	29482 (42,4 %)	24411 (89,1 %)	61269 (88,0 %)

TABLE 25 – Couverture de Wikipédia

La couverture de Wikipédia sur les entités nommées est correcte : moins de 15 % des distracteurs ne font pas référence à une page de Wikipédia. En revanche, la couverture de Wikipédia sur les chunks non entité nommée est moins large : plus d'un tiers des distracteurs ne font pas référence à une page de Wikipédia. Cela est dû au fait qu'une partie de ces chunks ne sont pas des entités, comme le terme «the little cat». Wikipédia étant principalement constitué d'entités, cette ressource ne couvre que des syntagmes nominaux.

En revanche, la grande majorité des réponses et des candidats des corpus évalués sont présents dans le corps des pages Wikipédia. La mesure fondée sur l'ESA peut donc s'effectuer sur une grande partie des entités nommées et des chunks non entité nommée. La mesure fondée sur l'ESA se calcule sur tout type de texte, elle peut donc se calculer sur tout type de chunk.

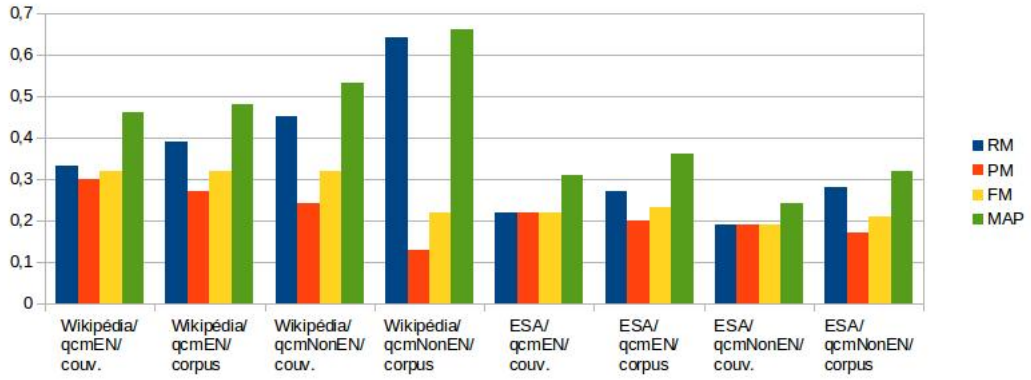


FIGURE 26 – Comparaison des résultats de l'évaluation des mesures fondées sur les corpus

4.4.3.2 *Calcul des mesures d'évaluation*

Le tableau 25 et la figure 26 montrent les résultats de l'évaluation des mesures fondées sur les corpus.

	Liens Wikipédia				ESA			
	qcmEN		**qcmNonEN**		**qcmEN**		**qcmNonEN**	
	Couv.	**Corpus**	**Couv.**	**Corpus**	**Couv.**	**Corpus**	**Couv.**	**Corpus**
RM	0,33	0,39	0,45	0,64	0,22	0,27	0,19	0,28
PM	0,30	0,27	0,24	0,13	0,22	0,20	0,19	0,17
FM	0,32	0,32	0,32	0,22	0,22	0,23	0,19	0,21
MAP	0,46	0,48	0,53	0,66	0,31	0,36	0,24	0,32

TABLE 26 – Résultats de l'évaluation des mesures fondées sur les corpus

Concernant les entités nommées, la mesure fondée sur les liens de Wikipédia est moins performante que les mesures fondées sur WordNet. La raison est que ces dernières mesures s'appuient sur la hiérarchie des concepts de WordNet, tandis que Wikipédia n'est pas une ressource hiérarchique. Cependant, la couverture de Wikipédia est considérablement plus large que celle de WordNet.

Les observations des chunks non entité nommée sont différentes : la mesure fondée sur les liens de Wikipédia est plus performante que les mesures fondées sur WordNet car la comparaison se fait au niveau des termes (et non au niveau des têtes syntaxiques

comme dans WordNet). Cependant, la couverture de Wikipédia est moins large que celle de WordNet.

Concernant la mesure fondée sur l'ESA, l'évaluation montre qu'elle est moins performante que la mesure fondée sur les liens Wikipédia. Cependant, sa couverture est plus large, étant donné qu'elle ne se restreint pas aux entités, mais à tout n-gramme dont les mots sont présents dans le corps des pages Wikipédia.

Nous observons également que la performance des mesures fondées sur les corpus sont similaires, quel que soit le type des options.

4.4.3.3 *Répartition des candidats selon le score fondé sur les mesures fondées sur les corpus*

Les figures 25k, 25l, 25m et 25n montrent la répartition des candidats selon leurs scores de voisinage sémantique fondés sur les corpus.

Ces figures montrent qu'en moyenne, les scores fondés sur les liens Wikipédia des distracteurs sont légèrement supérieurs à ceux des non-distracteurs. Cependant, une grande partie des candidats de type chunk non entité nommée n'ont pas de relation de voisinage sémantique avec la réponse (score nul).

En revanche, nous n'observons pas de distinction entre les scores fondés sur l'ESA des distracteurs et des non-distracteurs. De plus, la plupart des candidats ont un score très faible (inférieur à 0,02).

L'évaluation de ces mesures montre que les mesures fondées sur les corpus sont moins performantes que les mesures fondées sur les connaissances pour des termes de type entité nommée, contrairement aux chunks non entité nommée dont la performance est supérieure. Cependant, la couverture de Wikipédia sur ces chunks est limitée. La mesure fondée sur l'ESA complète cette couverture car sa couverture est large, quel que soit le type de termes.

Les remarques que nous avons donné pour chacune des mesures proposées sont synthétisées dans le tableau 27.

	Types		WordNet				Corpus	
	EN	DBpédia	reg	lch	jcn	lin	Wikipédia	ESA
Couverture (EN)	✓	✓	✗	✗	✗	✗	✓	✓
Couverture (non EN)	✗	✗	✓	✓	✓	✓	✓	✓
Relations	✓	✓	✓	✓	✓	✓	✓	✗
Hiérarchie	✓	✓	✓	✓	✓	✓	✗	✗
Corpus	✗	✗	✗	✗	✓	✓	✓	✓

TABLE 27 – Récapitulatif des propriétés des mesures proposées

4.5 MODÈLE D'ORDONNANCEMENT

Le classement des candidats selon les différents critères d'homogénéité que nous avons défini dans les sections précédentes est effectué avec SVMRank [3], un outil d'ordonnancement automatique par apprentissage supervisé fondé sur un modèle SVM (*Séparateur à Vaste Marge* ou *Support Vector Machine*). Un SVM est un classifieur discriminant défini par un hyperplan séparant les données des différentes classes. L'outil SVMRank compare les couples de distracteurs-non-distracteurs d'un même item et apprend le poids des critères tels que pour chaque couple de distracteur-non-distracteur (d, nd), $svm(d) > svm(nd)$, où $svm(c)$ est le score attribué au candidat c à partir de la combinaison des critères et des poids de chacun de ces critères, appris par SVM. Afin d'optimiser la performance du modèle, nous avons fait varier le paramètre C de SVMRank. Ce paramètre définit le compromis entre les erreurs d'apprentissage et la taille de la marge. Si la valeur de C est trop petite, la marge de l'hyperplan sera grande, ce qui peut impliquer un grand nombre d'erreurs d'apprentissage. À l'inverse, si la valeur de C est trop grande, la marge de l'hyperplan sera minimale, ce qui peut impliquer un surapprentissage. Nous avons également évalué le modèle avec différentes fonction de noyau : la fonction linéaire, et la fonction polynomiale. Les valeurs de C et le choix de la fonction noyau n'ont pas influé considérablement sur les résultats de l'évaluation du modèle, nous avons donc choisi les valeurs par défaut (C$= 0,01$ et fonction linéaire).

Pour évaluer notre modèle, nous avons constitué quatre modèles d'apprentissage, un par type de QCM (qcmEN et qcmNonEN) et par type d'évaluation (evalNDdocument et evalNDoption). Les attributs des modèles sont les mesures de voisinage sémantique que nous avons présentées dans ce chapitre. La plupart des attributs calculés sont communs aux quatre modèles :

1. la similarité des types sémantiques provenant de DBpédia (section 4.2.2) ;
2. le score de recoupement étendu de gloses (section 4.3) ;
3. le score de Leacock et Chodorow (section 4.3) ;
4. le score de Jiang et Conrath (section 4.3) ;
5. le score de Lin (section 4.3) ;
6. le score de comparaison des liens de pages de Wikipédia (section 4.4.1) ;
7. le score fondé sur l'ESA (section 4.4.2).

Les modèles calculés pour qcmEN, quel que soit l'évaluation, possèdent un attribut supplémentaire : la similarité des types d'entité nommée.

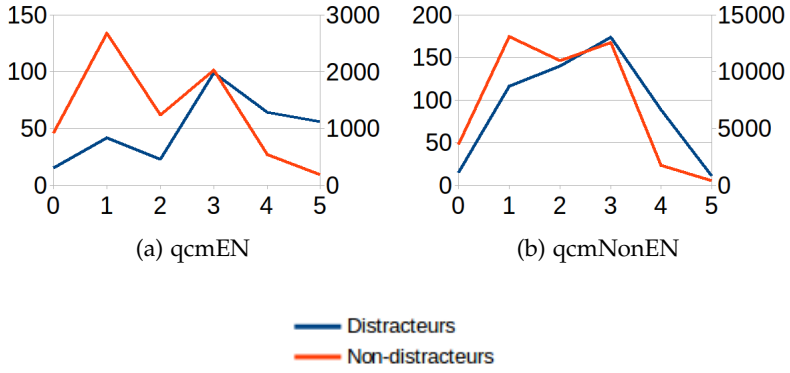
Les modèles calculés pour l'évaluation evalNDoption, quel que soit le type de QCM, possèdent un attribut supplémentaire : un attribut indiquant si le candidat apparaît dans le document de référence de l'item (1 si c'est le cas, 0 sinon).

3. http://www.cs.cornell.edu/people/tj/svmlight/svmrank.html

	qcmEN		qcmNonEN	
	R-D	**R-ND**	**R-D**	**R-ND**
DBpédia	69,6 %	32,8 %	5,3 %	3,1 %
WordNet	43,1 %	16,1 %	69,4 %	59,7 %
Corpus de fréquences	23,4 %	6,8 %	36,5 %	34,1 %
Wikipédia	80,9 %	50,3 %	45,7 %	14,1 %
ESA	91,0 %	82,6 %	87,0 %	82,0 %

TABLE 28 – Couvertures des ressources

Certaines ressources peuvent couvrir un petit nombre de réponses et de candidats, comme le montre le tableau 28, qui montre la proportion de couples de réponses-distracteurs (R-D) et de couples de réponses-non-distracteurs (R-ND) présents dans les ressources associées aux critères du modèle. Ce tableau montre que des ressources comme WordNet et Wikipédia sont faibles selon le type de termes : WordNet a une couverture faible sur les entités nommées et Wikipédia a une couverture faible sur les chunks non entités nommées.



(a) qcmEN (b) qcmNonEN

FIGURE 27 – Répartition des candidats en fonction du nombre de ressources les couvrant. Les abscisses représentent le nombre de ressources couvertes et les ordonnées représentent le nombre de distracteurs du corpus (à gauche des courbes) et le nombre de non-distracteurs du corpus (à droite des courbes)

La figure 27 montre qu'une grande partie des distracteurs sont couverts par trois ressources. Les entités nommées sont généralement couvertes par les types DBpédia, Wikipédia et les corps des pages Wikipédia tandis que les chunks non entités nommées sont généralement couvertes par WordNet, le corpus de fréquences associé à WordNet et les

corps des pages Wikipédia. Cette figure montre également que les non-distracteurs sont généralement couverts par moins de ressources que les distracteurs (ce qui corrobore les informations du tableau 28).

4.5.1 *Évaluation*

Nous évaluons le modèle d'ordonnancement sur chacun des corpus (qcmEN et qcmNonEN) et pour chacune des évaluations (evalNDdocument et evalNDoption) par une validation croisée en 7 sous-ensembles, c'est-à-dire que chacun des sous-ensembles du corpus est évalué selon le modèle appris à partir des autres sous-ensembles du corpus.

4.5.1.1 *Calcul des mesures d'évaluation*

	qcmEN				**qcmNonEN**			
	RM	**PM**	**FM**	**MAP**	**RM**	**PM**	**FM**	**MAP**
meme_type_EN	0,83	0,13	0,23	0,85				
wup (types DBpédia)	0,68	0,30	0,41	0,73	0,93	0,07	0,13	0,92
reg	0,68	0,22	0,34	0,73	0,44	0,14	0,21	0,46
lch	0,70	0,21	0,33	0,75	0,45	0,13	0,20	0,48
jcn	0,80	0,16	0,26	0,82	0,54	0,12	0,19	0,57
lin	0,81	0,17	0,28	0,83	0,55	0,11	0,18	0,57
Liens Wikipédia	0,39	0,27	0,32	0,48	0,64	0,13	0,22	0,66
ESA	0,27	0,20	0,23	0,36	0,28	0,17	0,21	0,32
Modèle	0,42	**0,40**	**0,41**	0,49	0,22	**0,22**	0,22	0,27

TABLE 29 – Résultats de l'évaluation des mesures et du modèle d'ordonnancement pour l'évaluation evalNDdocument

Dans le cas de l'évaluation evalNDdocument, le tableau 29 montre que le modèle d'ordonnancement obtient un meilleur équilibre entre le rappel et la précision que les mesures individuelles, quel que soit le corpus. Le modèle donne une meilleure précision que les autres mesures et de meilleurs résultats que les mesures fondées sur WordNet, utilisées par Mitkov *et al.* (2009).

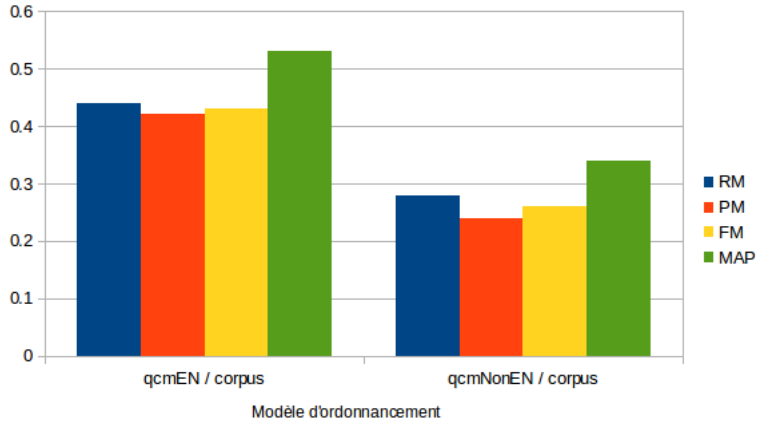


FIGURE 28 – Comparaison des résultats de l'évaluation du modèle d'ordonnancement pour l'évaluation evalNDdocument

	qcmEN	**qcmNonEN**
meme_type_EN	1,00	
wup (types DBpédia)	1,62	0,07
reg	0,10	0,03
lch	1,33	0,52
jcn	-0,09	0,08
lin	1,09	1,73
Liens Wikipédia	0,44	2,12
ESA	0,90	2,13

TABLE 30 – Poids des mesures appris dans le modèle pour l'évaluation evalNDdocument

Certaines mesures évaluées donnent un meilleur rappel que le modèle d'ordonnancement. Nous distinguons deux cas : le premier concerne les mesures fondées sur les types (d'entité nommée et spécifiques) qui sont plus efficaces pour filtrer les candidats que pour sélectionner les distracteurs. Le second cas concerne les mesures dont la couverture des ressources est faible (WordNet dans le corpus qcmEN et Wikipédia dans le corpus qcmNonEN). En effet, de telles mesures ne peuvent pas ordonner des candidats d'un item s'ils ne sont pas référencés par la ressource associée à la mesure. Le rappel indique la présence des distracteurs dans les premiers rangs parmi les candidats. Pour de tels items, le rappel est donc maximal ($rappel = 1$).

Les mesures donnent globalement des résultats inférieurs dans le corpus qcmNonEN. La raison principale est que les candidats et les réponses qui ne sont pas des entités nommées sont associés à moins d'informations sémantiques que les entités nommées, particulièrement sur les types sémantiques.

Dans le corpus qcmEN, la plupart des cas où les non-distracteurs ont un meilleur rang que les distracteurs sont dus au fait que les distracteurs et la réponse ne sont pas types par un type (de DBpédia) très spécifique. Parmi les non-distracteurs restants, ceux-ci sont assez pertinents pour être des distracteurs ou sont similaires à la réponse, donc ne peuvent être des distracteurs.

La majorité des non-distracteurs du corpus qcmNonEN de meilleur rang que les distracteurs sont clairement des non-distracteurs mais certaines mesures (particulièrement celles fondées sur WordNet) considèrent que ces non-distracteurs sont sémantiquement plus voisins que les distracteurs. Parmi les non-distracteurs restants, certains d'entre eux ne sont pas sémantiquement proches de la réponse dans le contexte courant (document de référence) ou sont assez pertinents pour remplacer les distracteurs.

Le tableau 30 montre que les mesures les plus importantes pour reconnaître les distracteurs de type entité nommée sont les mesures fondées sur les types, en particulier les types spécifiques. Pour les chunks non entité nommée, les mesures les plus importantes sont les mesures fondées sur les corpus. La mesure de Lin, fondée sur WordNet et sur les fréquences de mots dans un corpus, est également de poids fort. En revanche, la mesure fondée sur DBpédia a un poids négligeable pour reconnaître les distracteurs de type chunk non entité nommée. Cela peut s'expliquer par la très faible couverture de DBpédia sur ce type de candidats. Par ailleurs, les mesures de Jiang et Conrath et de recoupement automatique de gloses sont de poids négligeable, quel que soit le type de candidat appris.

Dans le cas de l'évaluation evalNDoption, le tableau 31 montre que les mesures individuelles donnent une précision plus faible que pour l'évaluation evalNDdocument. Cela est dû à deux causes principales. Premièrement, cette évaluation extrait plus de non-distracteurs que l'évaluation evalNDdocument, donc les mesures de faible couverture et/ou fondées sur les types donnent un très fort rappel et une très faible précision. Deuxièmement, un grand nombre de non-distracteurs sont sémantiquement plus proches de la réponse que les distracteurs, mais n'ont pas été sélectionnés manuellement car ils

	qcmEN				qcmNonEN			
	RM	**PM**	**FM**	**MAP**	**RM**	**PM**	**FM**	**MAP**
meme_type_EN	0,83	0,03	0,05	0,84				
wup (types DBpédia)	0,51	0,08	0,13	0,53	0,87	0,04	0,07	0,87
reg	0,60	0,12	0,20	0,62	0,42	0,11	0,18	0,43
lch	0,65	0,09	0,16	0,67	0,45	0,10	0,17	0,46
jcn	0,74	0,06	0,11	0,75	0,53	0,10	0,16	0,54
lin	0,74	0,07	0,12	0,76	0,53	0,09	0,16	0,55
Liens Wikipédia	0,35	0,25	0,29	0,39	0,58	0,13	0,21	0,60
ESA	0,28	0,21	0,24	0,33	0,30	0,19	0,23	0,33
Modèle	0,43	**0,42**	**0,43**	0,42	0,21	0,18	0,19	0,27

TABLE 31 – Résultats de l'évaluation des mesures et du modèle d'ordonnancement pour l'évaluation evalNDoption

n'apparaissent pas dans le contexte de l'item, soit le document de référence. En revenche, quelle que soit l'évaluation, le modèle d'ordonnancement donne les mêmes résultats pour le corpus qcmEN, contrairement au corpus qcmNonEN où l'évaluation evalNDdocument donne de meilleurs résultats.

Les résultats montrent que l'appartenance au document de référence est un critère important pour ordonner les entités nommées, contrairement aux chunks non entité nommée. En effet, un grand nombre de candidats sont des mots ou des n-grammes «communs» qui se retrouvent dans plusieurs documents comme le mot «track», ce qui fait que le critère d'appartenance à un document n'améliore pas le modèle d'apprentissage au niveau des chunks non entité nommée.

Nous avons également évalué la performance du filtrage des candidats similaires aux options, pour l'évaluation evalNDdocument. Pour cela, nous comparons les résultats de l'évaluation du modèle avec et sans les candidats filtrés.

	qcmEN	**qcmNonEN**
Sans filtrage	7583	42396
Avec filtrage	7700	42790

TABLE 32 – Nombre de non-distracteurs avec et sans filtrage des non-distracteurs similaires aux options

Le tableau 32 montre que le filtrage des non-distracteurs similaires aux options concerne peu de non-distracteurs (117 non-distracteurs dans le corpus qcmEN, 394 dans le corpus qcmNonEN, soit 1 % des non-distracteurs).

Pour les entités nommées, les non-distracteurs similaires aux options sont principalement des lieux liés par des relations de méronymie. Cependant, la ressource WordNet a une couverture assez faible sur les entités nommées. Une partie des non-distracteurs similaires aux options ne sont donc pas reconnues comme telles, à l'instar du non-distracteur «MacMurdo» (un centre de recherche situé en Antarctique) et de la réponse «Antarctica». Pour améliorer la reconnaissance de la similarité, il serait intéressant de prendre en considération les relations sémantiques de DBpédia indiquant une relation de similarité entre des entités.

Pour les chunks non entités nommées, un grand nombre de réponses et de candidats sont des termes dont seule la tête syntaxique réfère à des synsets de WordNet. La méthode de reconnaissance de termes similaires que nous avons proposée est inefficace sur ces termes. Pour ces cas, il serait intéressant d'effectuer une analyse sémantique plus fine afin de reconnaître d'éventuelles relations de similarité.

Le modèle donne des résultats similaires avec ou sans le filtrage des non-distracteurs similaires aux distracteurs. Cela peut s'expliquer par le fait qu'il existe peu de non-distracteurs reconnus comme similaires, et que cela n'influence pas l'évaluation. Une seconde possibilité serait que le modèle apprend les non-distracteurs similaires (reconnus ou non par notre méthode d'identification des non-distracteurs similaires). Ceux-ci se situeraient donc dans des rangs inférieurs aux distracteurs dans le classement.

Dans le chapitre 3, nous avons vu que la majorité des options du corpus qcmNonEN sont des syntagmes nominaux. Contrairement aux autres types de chunk, les syntagmes nominaux sont couvertes par toutes les ressources associées aux critères du modèle, même si la couverture de DBpédia et Wikipédia reste faible. En effet, les chunks verbaux, adjectivaux et adverbiaux ne sont pas disponibles dans DBpédia et Wikipédia car il ne s'agit pas d'entités. De plus, les chunks adjectivaux et adverbiaux ne sont pas traités pour trois des quatre mesures fondées sur WordNet. Seule la mesure de recoupement automatique de gloses couvre ces types de chunk.

Nous avons donc appris et évalué le modèle sur les syntagmes nominaux du corpus qcmChunk pour vérifier si sa performance est meilleure sur un type de chunk dont l'accès à toutes les ressources est possible. L'évaluation a été effectuée avec les non-distracteurs extraits dans le cadre de l'évaluation evalNDdocument. L'évaluation du modèle sur les syntagmes nominaux donne des résultats similaires au modèle évalué sur tous les chunks non entité nommée (qcmNonEN). Une raison possible est que les syntagmes nominaux représentent la majorité des chunks non entité nommée du corpus (environ 75 % du corpus qcmNonEN). Une autre raison est que l'absence de couverture des ressources (DBpédia, Wikipédia et WordNet) sur les autres chunks n'influence pas le calcul des poids des mesures dans le modèle.

4.6 CONCLUSION

Dans ce chapitre, nous avons présenté une méthode d'estimation de l'homogénéité appliquée à l'évaluation automatique de la qualité des distracteurs.

Cette méthode est fondée sur la combinaison par apprentissage de critères d'homogénéité sémantique. Dans le cadre d'application à l'évaluation automatique de la qualité des distracteurs, nous obtenons des résultats supérieurs aux méthodes de l'état de l'art. Un critère d'homogénéité syntaxique permet tout d'abord de filtrer les non-distracteurs non homogènes à la réponse. Les critères d'homogénéité sémantique sont représentés sous forme de mesures de voisinage sémantique. Contrairement aux travaux de l'état de l'art, nous avons pris en considération le type sémantique des termes. Pour cela, nous avons introduit une mesure calculant la similarité des types d'entité nommée des termes, et une mesure calculant la similarité de leurs types DBpédia selon la proximité de ceux-ci dans la taxonomie de DBpédia. Les mesures fondées sur la similarité des liens Wikipédia et l'ESA sont également des mesures qui n'ont pas été évaluées dans les travaux existants, même si une partie d'entre eux s'appuient sur des mesures de voisinage distributionnel. Parmi les mesures que nous avons évaluées, seules les mesures fondées sur WordNet ont été utilisées.

Nous avons étudié chacune des mesures du modèle pour observer leur comportement, et donc l'apport qu'elles offrent au modèle d'ordonnancement. Les mesures fondées sur la similarité du type des options permettent de donner une indication sur la similarité de leurs catégories sémantiques et les mesures fondées sur les relations sémantiques entre les termes ainsi que les mesures fondées sur les corpus permettent d'affiner la reconnaissance de l'homogénéité sémantique.

5 CONCLUSION ET PERSPECTIVES

5.1 CONCLUSION

Lors de cette thèse, nous avons travaillé dans le cadre de l'évaluation automatique de la qualité des distracteurs de QCM pédagogiques. Cela nous a amené à traiter de la problématique de l'estimation de l'homogénéité de deux termes. En effet, les études de QCM ont montré que l'homogénéité des options est un critère important pour rédiger des distracteurs pertinents.

Pour répondre à la problématique de l'homogénéité, nous avons cherché à répondre aux trois questions sous-tendant cette problématique :

— Comment définir l'homogénéité syntaxique et sémantique dans un but de reconnaissance automatique ?

— Quelles tâches de Traitement Automatique des Langues utiliser pour reconnaître l'homogénéité ?

— Quel modèle permet d'estimer globalement l'homogénéité ?

Nous avons défini l'homogénéité en nous appuyant sur un corpus de QCM dont les distracteurs ont été annotés selon différents degrés d'homogénéité syntaxique et sémantique. Ces annotations ont permis de valider la définition de l'homogénéité et de vérifier la possibilité de reconnaître l'homogénéité par des méthodes automatiques. Cette vérification s'est appuyée sur des méthodes d'analyse syntaxique et d'annotation en entités nommées des options. L'évaluation de ces méthodes a montré qu'il est possible de reconnaître automatiquement l'homogénéité des options.

Cela nous a conduit à développer un modèle d'estimation de l'homogénéité appliqué à l'évaluation automatique de la qualité des distracteurs. Ce modèle est un modèle d'ordonnancement des candidats et est capable de reconnaître le degré de pertinence des distracteurs selon leurs rangs parmi les candidats. Les critères d'homogénéité du modèle s'appuient sur une comparaison des arbres syntaxiques des candidats et de la réponse, et de mesures de voisinage sémantique fondées sur des connaissances sémantiques structurées et sur de vastes corpus de documents. Notre modèle a été conçu pour des réponses de type entité nommée et de type chunk non entité nommée, ainsi que des syntagmes nominaux. Nous avons proposé un mode d'évaluation automatique de l'homogénéité sur un corpus de QCM, ce qui permet de comparer différentes méthodes sur une même base.

5.2 PERSPECTIVES

Dans cette section, nous présentons quelques perspectives s'inscrivant dans la continuité de nos travaux de thèse.

Une première perspective consiste à intégrer notre modèle dans des systèmes d'aide à la rédaction de QCM et de sélection automatique de distracteurs. Cette dernière intégration serait particulièrement intéressante car elle permettrait de comparer notre modèle aux méthodes de sélection automatique de distracteurs proposées dans l'état de l'art. Cette comparaison nous amènerait à évaluer les distracteurs sélectionnés sur des apprenants selon des critères pédadogiques. En effet, la pertinence des distracteurs serait mesurée selon les options sélectionnées par les apprenants. Dans le chapitre 2, nous avons vu que ce type d'évaluation s'effectue avec des mesures psychométriques, estimant ainsi le niveau de fiabilité, de validité et de pertinence des distracteurs sélectionnés.

Une autre perspective consiste à améliorer la reconnaissance de la similarité des termes. En effet, notre méthode de reconnaissance de la similarité s'appuie principalement sur les relations sémantiques de WordNet. Elle reconnaît également des termes similaires s'ils font référence à la même entité DBpédia. Cependant, les expériences nous ont montré qu'une partie des termes similaires ne sont pas reconnus. Concernant les entités nommées, cela est dû à la faible couverture de WordNet. La prise en considération des relations sémantiques de DBpédia pourrait améliorer la reconnaissance de la similarité des entités nommées. Nous pourrions faire une étude de ces relations pour identifier celles qui indiquent des relations de similarité (en l'ocurrence, des relations de méronymie). Concernant les chunks non entités nommées, les termes similaires non reconnus par la méthode que nous avons proposée sont des n-grammes dont seule la tête fait référence à un synset de WordNet. Il serait intéressant de proposer des analyses syntaxiques et sémantiques plus fines pour identifier les relations de similarité entre de tels termes.

Dans la continuité de notre travail, il serait intéressant d'étudier la pertinence d'un modèle intégrant des critères d'homogénéité syntaxique et sémantique. En effet, nous nous sommes focalisé sur l'homogénéité sémantique des termes mais la prise en considération de critères d'homogénéité syntaxique pourrait donner une indication différente sur le degré d'homogénéité des termes comparés.

Il serait également intéressant d'adapter notre modèle pour d'autres langues, en particulier le français. Les critères d'homogénéité que nous avons proposés s'appliquent à des langues différentes si les ressources existent. Nous avons évalué quelques critères d'homogénéité sur un corpus de QCM en français, et les résultats ont montré que la performance de ces critères est la même, quelle que soit la langue. Cependant, certaines ressources, comme des bases lexicales structurées, sont moins riches pour la langue française.

Enfin, il serait intéressant d'adapter notre modèle d'ordonnancement à tous types d'options. Cela nécessiterait de définir la notion d'homogénéité syntaxique d'options longues.

Première partie

ANNEXE

A

FORMATS D'ITEMS À CHOIX MULTIPLES

Il existe plusieurs formats d'items à choix multiples. Chacun de ces formats a des avantages et des inconvénients du point de vue de la compréhension de l'apprenant et de l'effort fourni par l'enseignant pour les rédiger. D'après Haladyna *et al.* (2002), les sept formats les plus fréquents sont les suivants : les choix multiples conventionnels (*Conventional MC*), les choix alternatifs (*Alternate-Choice*), les vrai-faux (*True-False*), les vrai-faux multiples (*Multiple True-False*), les correspondances (*Matching*), les choix multiples complexes (*Complex MC*) et les ensembles d'items dépendant d'un contexte (*Context-dependant Item Set*). Ces formats sont fondés sur la forme de l'amorce ou des options, comme le montre le tableau 33.

Amorce	+ scénario	items dépendant d'un contexte
Options	2 op.	choix alternatifs, vrai-faux, multiples vrai-faux
	+ de 2 op.	choix multiples conventionnel, correspondances, choix multiples complexes

TABLE 33 – Formats des items classés selon qu'ils sont fondés sur la forme de l'amorce ou des options

Dans cette section, nous présentons chacun de ces formats d'items.

A.1 CHOIX MULTIPLES CONVENTIONNELS

Ce format d'item est le plus couramment utilisé dans les QCM. Il existe deux types de variations de ce type d'items, fondées sur la forme de l'amorce : celle-ci est une interrogation ou un texte à trous (*Fill-In The Blank Item (FBI)*) à compléter par les options. Au niveau de la qualité pédagogique de ces types d'items, il n'existe pas de grande différence, bien que Haladyna *et al.* (2002) ont une préférence pour les items dont l'amorce est une interrogation car ils considèrent qu'ils sont plus directs pour énoncer l'idée centrale de l'item. L'exemple ci-dessous montre un item à choix multiples conventionnel.

Which of the following most clearly defines the process of pollination ?

A The joining of egg and sperm cells.

B The transfer of pollen grains to the pencil.

C Food is broken down and energy is released.

A.2 CHOIX ALTERNATIFS

Les items à choix alternatif sont des items contenant deux options : la réponse et un distracteur. Les avantages de ce type d'items sont qu'ils sont plus faciles à rédiger que les items conventionnels (au niveau des distracteurs, étant donné qu'il n'y en a qu'un à engendrer). De plus, pour les évaluations d'items conventionnels, les apprenants doués ont tendance à restreindre le nombre d'options à deux (éliminant ainsi les distracteurs supposés), ces options étant pour eux les plus plausibles. L'exemple ci-dessous montre un item à choix alternatifs.

> Which of the following would most effectively slow down the process of respiration in plants ?
>
> A Cold weather
>
> B Stormy weather

A.3 VRAI-FAUX

Les items «vrai-faux» sont des items à choix alternatif dont les options sont des choix binaires tels que vrai/faux, oui/non, correct/incorrect, fait/opinion... L'exemple ci-dessous montre un item «vrai-faux».

> The capital of Uruguay is Montevideo.
>
> A True.
>
> B False.

A.4 MULTIPLES VRAI-FAUX

Le format des items à multiples vrai-faux est une combinaison entre celui des items à choix multiples conventionnels et celui des items «vrai-faux». En effet, à partir d'un item principal ou d'un scénario, les apprenants évaluent chacune des options en indiquant si elles sont vraies ou fausses. De tels items sont plus faciles à rédiger que les items conventionnels. L'exemple ci-dessous montre un exemple d'item à multiples vrai-faux.

> You are an expert organic farmer. You know the secrets of growing strong, healthy plants. Which of the following would describe your farming practices ? (Mark A if true, B if false.)

A.5 CORRESPONDANCES

Les items à correspondances nécessitent un ensemble d'options suivies par un ensemble d'amorces correspondantes (affirmations, questions ou syntagmes). Cependant, ce format est rarement utilisé. L'exemple ci-dessous montre un exemple d'item à correspondances.

1. When you plant some beans you make certain that the beans will be well shaded to receive little to no light.	A	B
2. When you plant your seeds you make sure to water them them and continue to keep the soil moist.	A	B
3. You plant your seeds only when the temperature is appropriate.	A	B
4. Because you know how pollination occurs, you spray your crops with insecticides to prevent bees and other insects from harming your crops.	A	B

Match each term on the right with the description on the left.

1. Attracts bees
2. Produces pollen grains
3. Houses the eggs cells
4. Seeds are formed
5. Contains the ovary

A Pollen grains
B Petals
C Flower
D Stamen
E Ovary
F Pistil

A.6 CHOIX MULTIPLES COMPLEXES

Les items à choix multiples complexes contiennent une ou plusieurs réponses. L'apprenant doit identifier la (les) réponse(s) en sélectionnant la combinaison la (les) représentant. En psychologie de l'éducation, ce format d'item n'est pas conseillé du point de vue des enseignants et des apprenants : ces items nécessitent plus de temps de rédaction et biaisent la notation des apprenants car elles donnent des indices sur les réponses aux apprenants qui connaissent partiellement le domaine évalué en choisissant uniquement entre les options contenant les réponses qu'ils connaissent. L'exemple ci-dessous montre un item à choix multiples complexe.

Which of the following are fruits ?

1. Tomatoes
2. Tomatillos
3. Habaneros peppers

A 1 & 2

B 2 & 3

C 1 & 3

D 1, 2 & 3

A.7 ITEM DÉPENDANT D'UN CONTEXTE OU ENSEMBLE D'ITEMS

Les items dépendant d'un contexte comportent un contexte tel qu'un scénario, une vignette, un tableau, un graphique, un passage textuel ou un autre matériel visuel. Ce contexte est suivi de l'item. Ce format est très utilisé mais il prend un espace considérable dans un test et nécessite un temps d'administration plus élevé (pour que les apprenants puissent étudier le contexte). Une variante importante de ce format est l'ensemble d'items dépendant du même contexte, lui aussi fortement utilisé. L'exemple ci-dessous montre un ensemble d'items dépendant d'un contexte.

Imagine you are a delegate from Massachussets to the Constitutional Convention. You have been authorized to act on behalf of your state.

1. You would most likely approve of the
   A New Jersey Plan.
   B Virginia Plan.
2. You would oppose the three-fifths compromise because
   A Your state, as a rule, is strongly abolotionist.
   B You will be grossly outrepresented in Congress by northern states.
   C You want only a single representative house.
3. You support the suggestion that Congress tax
   A Imports.
   B Experts.
4. Because of your state's experience with Shays' Rebellion, you feel
   A Farmers shouls not have to carry the tax burden for townspeople.
   B Native Americans must be pacified before there can be peace.
   C Tories ought to pay reparations.

B

TAXONOMIE DE CONSIGNES DE RÉDACTION DE Haladyna *et al.* (2002)

CONTENT CONCERNS

1. Every item should reflect specific content and a single specific mental behavior, as called for in test specifications (two-way grid, test blueprint).
2. Base each item on important content to learn ; avoir trivial content.
3. Use novel material to test higher level learning. Paraphrase textbook language or language used during instruction when used in a test item to avoid testing for simply recall.
4. Keep the content of each item independant from content of other items on the test.
5. Avoid over-specific and over general content when writing MC items.
6. Avoid opinion-based items.
7. Avoid trick items.
8. Keep vocabulary simple for the group of students being tested.

FORMATTING CONCERNS

9. Use the question, completion, and best answer versions of the conventional MC, the alternate choice, true-false (TF), multiple true-false (MTF), matching, and the context-dependent item and item set format, but AVOID the complex MC (Type K) format.
10. Format the item vertically instead of horizontally.

STYLE CONCERNS

11. Edit and proof items.
12. Use correct grammar, punctuation, capitalization, and spelling.
13. Minimize the amount of reading in each item.

WRITING THE STEM

14. Ensure that the directions in the stem are very clear.

15. Include the central idea in the steam instead of the choices.
16. Avoid window dressing (excessive verbiage).
17. Word the stem positively, avoid negatives such as NOT or EXCEPT. If negative words are used, use the word cautiously and always ensure that the word appears capitalized and boldface.

WRITING THE CHOICES

18. Develop as many effective choices as you can, but research suggests three is adequate.
19. Make sure that only one of these choices is the right answer.)
20. Vary the location of the right answer according to the number of choices.
21. Place choices in logical or numerical order.
22. Keep choices independant ; choices should not be overlapping.
23. Keep choices homogeneous in content and grammatical structure.
24. Keep the length of choices about equal.
25. *None-of-the-above* should be used carefully.
26. Avoid *All-of-the-above*.
27. Phrase choices positively ; avoid negatives such as NOT.
28. Avoid giving clues to the right answer, such as :
    a) specific determiners including always, never, completely, and absolutely ;
    b) clang associations, choices identical to or resembling words in the stem ;
    c) grammatical inconsistencies that cue the test-taker to the correct choice ;
    d) conspicuous correct choice ;
    e) pairs or triplets of options that clue the test-taker to the correct choice ;
    f) blatantly absurd, ridiculous options.
29. Make all distractors plausible.
30. Use typical errors of students to write your distractors.
31. Use humor if it is compatible with the teacher and the learning environment.

BIBLIOGRAPHIE

Mohammed Elhassan Abdalla, Abdelrahim Mutwakel Gaffar et Rasha Ali Suliman : *Constructing A-Type Multiple Choice Questions (MCQs) : Step By Step Manual*. Abdelrahim Mutwakel Gaffar, 2011.

Ahmad A Abdel-Hameed, Eiad A Al-Faris, Ibrahim A Alorainy et Mohammed O Al-Rukban : The criteria and analysis of good multiple choice questions in a health professional setting. *Saudi medical journal*, 26(10):1505–1510, 2005.

Steven P Abney : *Parsing by chunks*. Springer, 1992.

Thibault André : Génération automatique de distracteurs dans le cadre de qcm, 2013.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak et Zachary Ives : Dbpedia : A nucleus for a web of open data. *In The semantic web*, pages 722–735. Springer, 2007.

Satanjeev Banerjee et Ted Pedersen : Extended gloss overlaps as a measure of semantic relatedness. *In IJCAI*, volume 3, pages 805–810, 2003.

Huguette Bernard et France Fontaine : *Les questions à choix multiple : guide pratique pour la rédaction, l'analyse et la correction*. Montréal : Service pédagogique, Université de Montréal, 1982.

Benjamin S Bloom : Taxonomy of educational objectives : Handbook i : Cognitive domain. *New York : David McKay*, 19:56, 1956.

Steven J Burton, Richard R Sudweeks, Paul F Merrill et Bud Wood : *How to prepare better multiple-choice test items : Guidelines for university faculty*. Brigham Young University Testing Services, 1991.

Rudi L Cilibrasi et Paul MB Vitanyi : The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):370–383, 2007.

Julie Considine, Mari Botti et Shane Thomas : Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian*, 12 (1):19–24, 2005.

Marija Cubric et Milorad Tosic : Towards automatic generation of e-assessment using semantic web technologies. *International Journal of e-Assessment*, 2011.

Ido DAGAN, Lillian LEE et Fernando PEREIRA : Similarity-based methods for word sense disambiguation. *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–63. Association for Computational Linguistics, 1997.

Joachim DAIBER, Max JAKOB, Chris HOKAMP et Pablo N MENDES : Improving efficiency and accuracy in multilingual entity extraction. *In Proceedings of the 9th International Conference on Semantic Systems*, pages 121–124. ACM, 2013.

Steven M DOWNING : The effects of violating standard item writing principles on tests and students : the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, 10(2):133–143, 2005.

Christiane FELLBAUM : Wordnet : An electronic database, 1998.

Olivier FERRET : Similarité sémantique et extraction de synonymes à partir de corpus. *Actes de TALN 2010 Traitement Automatique des Langues Naturelles-TALN 2010*, 2010.

Olivier FERRET : Sélection non supervisée de relations sémantiques pour améliorer un thésaurus distributionnel. *In 20eme Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013)*, pages 48–61, 2013.

Olivier FERRET, Brigitte GRAU, Martine HURAULT-PLANTET, Gabriel ILLOUZ, Christian JACQUEMIN, Nicolas MASSON et Paule LECUYER : Qalc–the question-answering system of limsi-cnrs. *In TREC*, 2000.

Jenny Rose FINKEL, Trond GRENAGER et Christopher MANNING : Incorporating non-local information into information extraction systems by gibbs sampling. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

John Rupert FIRTH, W HAAS, Michael AK HALLIDAY, WS ALLEN, RH ROBINS, FR PALMER, J CARNOCHAN et TF MITCHELL : *Studies in linguistic analysis*. Blackwell, 1962.

Evgeniy GABRILOVICH et Shaul MARKOVITCH : Computing semantic relatedness using wikipedia-based explicit semantic analysis. *In IJCAI*, volume 7, pages 1606–1611, 2007.

Thomas M HALADYNA et Steven M DOWNING : A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1):37–50, 1989.

Thomas M HALADYNA, Steven M DOWNING et Michael C RODRIGUEZ : A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3):309–333, 2002.

Jay J JIANG et David W CONRATH : Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.

Nikiforos Karamanis, Le An Ha et Ruslan Mitkov : Generating multiple-choice test items from medical text : A pilot study. *In Proceedings of the Fourth International Natural Language Generation Conference*, pages 111–113. Association for Computational Linguistics, 2006.

Dan Klein et Christopher D Manning : Accurate unlexicalized parsing. *In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics, 2003.

Peter Kolb : Disco : A multilingual database of distributionally similar words. *Proceedings of KONVENS-2008, Berlin*, 2008.

David R Krathwohl : A revision of bloom's taxonomy : An overview. *Theory into practice*, 41(4):212–218, 2002.

Thomas K Landauer et Susan T Dumais : A solution to plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

Claudia Leacock et Martin Chodorow : Combining local context and wordnet similarity for word sense identification. *WordNet : An electronic lexical database*, 49(2):265–283, 1998.

John Lee et Stephanie Seneff : Automatic generation of cloze items for prepositions. *In INTERSPEECH*, pages 2173–2176, 2007.

Roger Levy et Galen Andrew : Tregex and tsurgeon : tools for querying and manipulating tree data structures. *In Proceedings of the fifth international conference on Language Resources and Evaluation*, pages 2231–2234. Citeseer, 2006.

Anne-Laure Ligozat : Question classification transfer. *In ACL (2)*, pages 429–433, 2013.

Dekang Lin : Using syntactic dependency as local context to resolve word sense ambiguity. *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 64–71. Association for Computational Linguistics, 1997.

Leslie A Miller, Robert L Lovler et Sandra A McIntire : *Foundations of psychological testing : A practical approach*. Sage Publications, 2012.

David Milne et Ian H Witten : An open-source toolkit for mining wikipedia. *Artificial Intelligence*, 194:222–239, 2013.

Ruslan Mitkov, Le An Ha, Andrea Varga et Luz Rello : Semantic similarity of distractors in multiple-choice tests : extrinsic evaluation. *In Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 49–56. Association for Computational Linguistics, 2009.

A. Papasalouros, K. Kotis et K. Kanaris : Automatic generation of multiple-choice questions from domain ontologies. *IADIS e-Learning*, 2008.

Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Richard Sutcliffe et Roser Morante : Qa4mre 2011-2013 : Overview of question answering for machine reading evaluation. *In Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 303–320. Springer, 2013.

Van-Minh Pho, Thibault André, Anne-Laure Ligozat, B Grau, G Illouz et Thomas François : Multiple choice question corpus analysis for distractor characterization. *In 9th International Conference on Language Resources and Evaluation (LREC 2014)*, 2014.

Van-Minh Pho, Anne-Laure Ligozat et Brigitte Grau : Distractor quality evaluation in multiple choice questions. *In Artificial Intelligence in Education*, pages 377–386. Springer, 2015a.

Van-Minh Pho, Anne-Laure Ligozat et Brigitte Grau : Estimation de l'homogénéité sémantique pour les questionnaires à choix multiples. *Actes de TALN 2015 Traitement Automatique des Langues Naturelles-TALN 2015*, 2015b.

Simone Paolo Ponzetto et Michael Strube : Knowledge derived from wikipedia for computing semantic relatedness. *J. Artif. Intell. Res.(JAIR)*, 30:181–212, 2007.

Roy Rada, Hafedh Mili, Ellen Bicknell et Maria Blettner : Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on*, 19 (1):17–30, 1989.

Sheldon Ross : A first course in probability, 1976.

Jack Snowman : Educational psychology : What do we teach, what should we teach ? *Educational Psychology Review*, 9(2):151–170, 1997.

Marie Tarrant, Aimee Knierim, Sasha K Hayes et James Ware : The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education in Practice*, 6(6):354–363, 2006.

Marie Tarrant et James Ware : Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, 42(2): 198–206, 2008.

Peter Turney : Mining the web for synonyms : Pmi-ir versus lsa on toefl. 2001.

Zhibiao Wu et Martha Palmer : Verbs semantics and lexical selection. *In Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

Kaizhong ZHANG et Dennis SHASHA : Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing*, 18(6):1245–1262, 1989.